AD_____

Award Number: W81XWH-13-1-0020

TITLE: Health-Terrain: Visualizing Large Scale Health Data

PRINCIPAL INVESTIGATOR: Ph.D. Fang, Shiaofen

CONTRACTING ORGANIZATION: Indiana University

REPORT DATE: APR 14

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
April 2014	Annual	7 MAR 2013 – 6 MAR 2014

4. TITLE AND SUBTITLE
Health-Terrain: Visualizing Large Scale Health Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-13-1-0020

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shiaofen Fang, Mathew Palakal, Yuni Xia, Shaun J. Grannis, Jennifer L. Williams

email: sfang@cs.iupui.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Trustees of Indiana University

Indianapolis, IN 46202

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
In the past year, we have made significant progress including: (1) creating a concept space data model, which represents a schema tailored to support diverse visualizations and provides a uniform ontology that allows the system to be leveraged for many types of health care datasets through individually designed text and data mining procedures; (2) designing and implementing data and text mining analytics and visualization algorithms; and (3) developing a flexible prototype system for using our analytics and visualization framework to explore large-scale, real-world health data. These three components are integrated in a generalizable browser-based graphical interface, which enables flexible and free-form data exploration and hypothesis discovery and also a more flexible distribution of the resulting software. We have completed the majority of algorithm development and implementation; implementation and testing of a few remaining advanced visualization techniques are outstanding. The system has received favorable initial feedback from users, and we believe it has potential as an open source tool to support health data visualization tasks.

15. SUBJECT TERMS
Health Data, Visualization, large-scale

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	38	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

The goal of this project is to develop novel visualization techniques and tools for large and complex health care data to facilitate timely decision-making and trend/pattern detection. A prototype system will be developed to test the effectiveness of this approach on a large-scale health care database that is currently available at Regenstrief Institute. We will develop a public health use case leveraging a Notifiable Condition Detector (NCD) dataset that contains reportable disease conditions that are transmitted to Indiana public health authorities (over 800,000 reports). Clinicians and public health stakeholders seek to uncover informative trends contained within the growing population-based datasets. To support knowledge discovery, in this project, we first extract meaningful terms and their associations and attributes from the raw data by applying data mining and text mining algorithms to construct a concept space. A browser-based user interface will also be developed to enable interactive online data exploration. A suite of visualization algorithms and techniques will be developed and implemented within the prototype system. Visualizations include a novel 3D spatiotemporal terrain visualization technique for big time-series data over the Indiana geographical area.

Body

The three primary goals for the first year of the project include: (1) Concept space definition; (2) Algorithm design; and (3) System design and implementation. We have accomplished goals 1 and 2 and have made substantial progress on goal 3. We implemented multiple advanced visualization algorithms and expect to complete implementation of the remaining visualizations in approximately 2-4 weeks.

1. Concept Space Definition

The "concept space" represents a uniform layer of clinical observations and their associations and serves as a platform for users to explore the data using visualization and analysis methods. The concept terms are derived from data mining and text-mining processes applied to the use case datasets. For this project we focus on a population health use case that leverages an automated Notifiable Condition Detector (NCD). The NCD dataset contains 833,710 notifiable cases spanning more than 10 years from among 439,547 unique patients. An additional dataset linked to the original NCD patient's data was extracted from the Indiana Network for Patient Care (INPC) health information exchange containing 325,791 unstructured clinical discharge summaries, laboratory reports, and patient histories. Disease concepts were extracted from the NCD dataset. Text mining algorithms were then applied to additional linked text dataset (unstructured clinical summaries) to construct ontologies for different concept types, including Disease, Symptom, Mental behavior, and Risky Behavior. An association-mining algorithm was applied to the combined terms to generate an association graph among all the concepts terms. The resulting concept space, along with the processed NCD data, is represented in a data model designed to support our specific ontology.

Considering the visualization-specific requirements, we designed a three-layer data model (Fig.1) to store the NCD and supporting text dataset. The first layer contains base tables for the entities included in our ontology: patient, disease, location, and other terms. The table for these additional terms has four subcategories: mental behavior, risky behavior, medication and symptom. The second layer contains associations between the primary patient entity and additional three supporting entities. The third layer contains indirect associations between disease, term and location and was constructed using data mining techniques. Designs for the specific supporting schema and classes of associations were informed by the data mining results and the data elements necessary to support each specific visualization. Further, to avoid costly database scans during visualization execution the schema also includes pre-computed aggregate data necessary to support the specific visualizations. Pre-computed aggregate data include joint statistics such as entity association frequencies, e.g., the number of instances of "disease X" associated with "location Y".

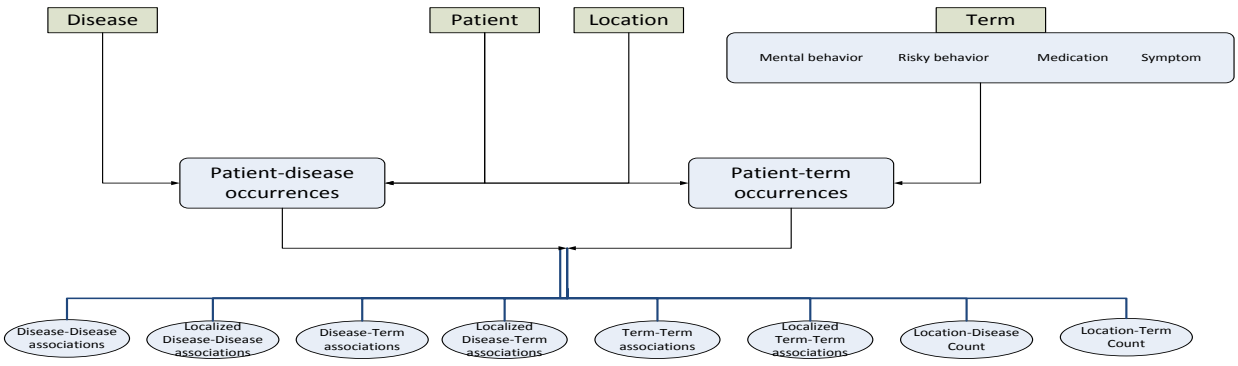


Figure 1: Database Model

To preserve patient confidentiality we created a randomly assigned, pseudonymized patient identifier, (called “PseudoID”) linking records within and among the NCD and INPC datasets, but conveys no identifiable traits. In rare instances a PseudoID may falsely match more than one patient. To avoid this error, we used three additional fields to verify that records sharing the same PseudoID represent the same patient. The three additional fields are gender, race, and date of birth. If two records have the same PseudoID but disagree on non-null genders with values of ‘M’ or ‘F’, then the records are treated as separate patients. The race validation rule functions similarly to gender validation. For date of birth validation, we standardize date of birth format as “yyyy-mm-dd” and apply the longest common subsequence string comparator[7]. If the ratio between the length of the longest common subsequence over the length of yyyy-mm-dd format is less than a certain threshold, date of birth validation fails. Records are determined to represent the same patient only if all three additional fields pass validation.

After cleansing, the database contained 439,547 patients, 1,976 diseases, 3,756 locations and 3,851 terms (711 symptoms, 93 risky behaviors, 200 mental behaviors and 2847 medications). The second layer of the database contains 1,302,173 disease occurrences and 1,215,659 term occurrences. At least 90,376 patients are associating with at least one term non-disease. All of these patients have a least one disease. The number of patients having more than one disease is 114,820, which is later used for association mining. At the third layer, the database contains 577,888 global associations between two different diseases, 1,958,227 global associations between two different terms and 1,032,864 global associations between a disease and a term. We removed duplicate public health case reports, which were defined as record having the same patient, the same date, and the same notifiable condition. We subsequently identified the most common reported conditions. The condition “Lead Exposure” was found among 256,823 patients. However, lead poisoning is not common in practice. “Lead Exposure” has the highest occurrence because Indiana’s reporting law requires that all laboratories performing blood lead tests report the results of those tests, whether normal or abnormal. Therefore, even when the test result is in the normal range, the test was reported. It leads to a high number of records on “Lead Exposure” in the data, while most of the report has negative results. Additional frequently reported conditions included: 1) Staphylococcus Methicillin-Resistant Aureus (MRSA), 2) HIV, 3) Chlamydia Infection, 4) Hepatitis B, 5) Hepatitis C, 6) Gonorrhea, 7) Chickenpox, 8) Measles, 9) Hepatitis A, 10) Enterococcus Vancomycin-Resistant (VRE) 11) Trichomoniasis 12) Syphilis. The figure (Fig. 2) below represents the occurrence rate of the most common diseases.

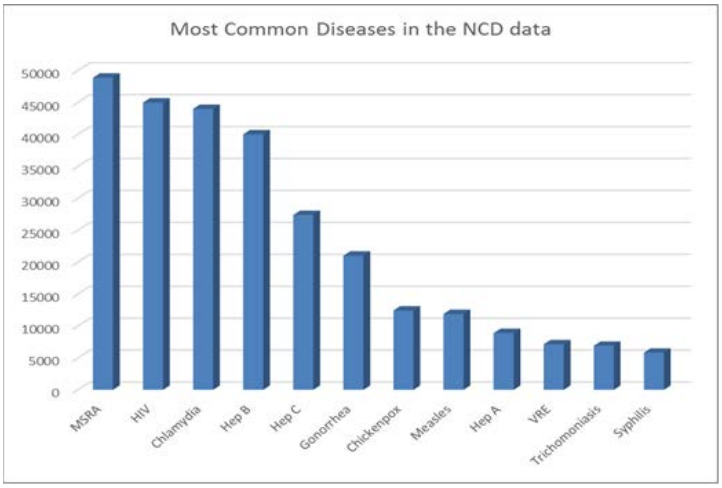


Figure 2: Most Common Diseases in the NCD dataset

We analyzed the disease distribution across races. In the figure (Fig. 3) below we compare the difference between the two largest races: white and black. The result is shown in Figure 3, with the black bar representing the occurrence percentage of each disease among black patients and the blue bar representing the occurrence percentage of disease among white patients. It shows that among black patients, Chlamydia Infection and Gonorrhea are the most common conditions in the NCD data. Trichomoniasis and Syphilis are also more common in black patients than in white patients. Among white patients, the most common condition is Staphylococcus Methicillin-Resistant Aureus (MSRA).

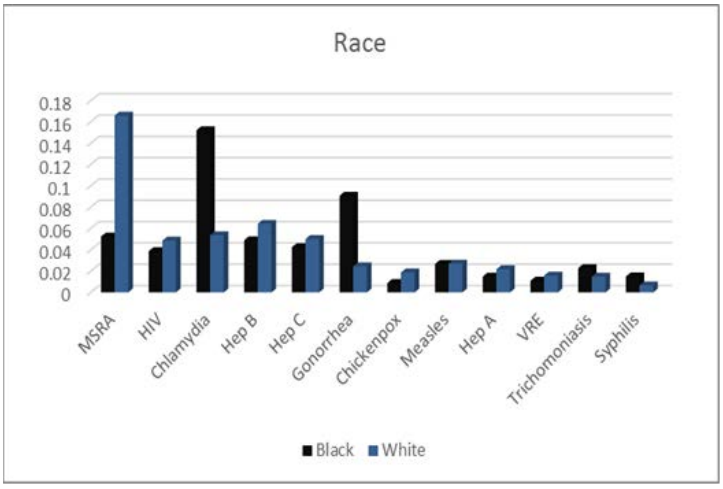


Figure 3: Diseases Distribution Across Races

We test the database efficiency by three sets of common samples queries designed by visualization and health science experts. The first query set is about geographical distribution of one or a combination of diseases. The second query set retrieves strong associated diseases to a given disease. The third query set finds common diseases occurring at a given range of age. Table 1 summarizes the performance of three types of queries and suggests that the further optimization will be required for effective user interactions during interactive visualization.

Query set	Example	Involved tables	Runtime
1	Geographical (at city level) distribution of chlamydia	location, diseases, patient-disease occurrences	12s
2	List the diseases associating with chlamydia	diseases, associations	0.5s
3	What are the most common diseases for patient age from 20 to 40	diseases, patients, patient-disease occurrences	16s

Table 1: 3 query set for testing database

2. Algorithm Development

There are 3 types of algorithms that need to be developed: (1) Text Mining algorithms to extract concept terms from clinical notes; (2) Data mining algorithms, such as association mining and clustering; and (3) visualization algorithms for various visualization tools.

a. Text Mining

The NCD receives clinical data that includes the diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations. But much of the information pertaining the patients' condition is available in the clinical reports. Mining these reports can provide a bigger picture of various other conditions that the patient experienced during his/her treatment. This information can provide valuable insights on the patients' socioeconomic condition, behavior risk factors, environmental factors and genetic information (family history). Natural Language Processing (NLP) provides a means to augment the NCD data analytics with the information discovered from these clinical reports.

NLP techniques were carried out to process 325791 clinical notes that contain patient discharge summaries, laboratory reports, patient history, etc. Although these records are de-identified due to which the patient specific information are absent, a pseudo-patient Id has been provided to help process the reports. Basic processing of the reports was performed for converting the clinical notes from XML format to simple text format and sentence splitting. Advanced level NLP was applied in the form of named entity recognition (NER) for extracting diseases, symptoms, mental behavior, risky behavior and medication information from the reports. This was done with the help of UMLS [7] database which is a repository of clinical and health related terms. Once the entities were extracted using NER, negation analysis was applied using NEGEX algorithm [8] to remove negated terms. Fig. 4 shows the process that was used in extracting the vital information from the reports.

The advantage of using UMLS is that all variations of clinical terms get captured that provide a large set of terms available for further analysis. For example, clinical notes that indicate "Hepatitis" contains terms like "Hepatitis", "Hepatitis B", "Hep", "Hep B" etc. The large number of terms extracted contains different occurrences of the same diseases, symptoms, etc. We apply stemming and grouping algorithms to group these terms to reduce the total number of terms. The identified terms are stored in different data tables and joined using the pseudo-patient Id.

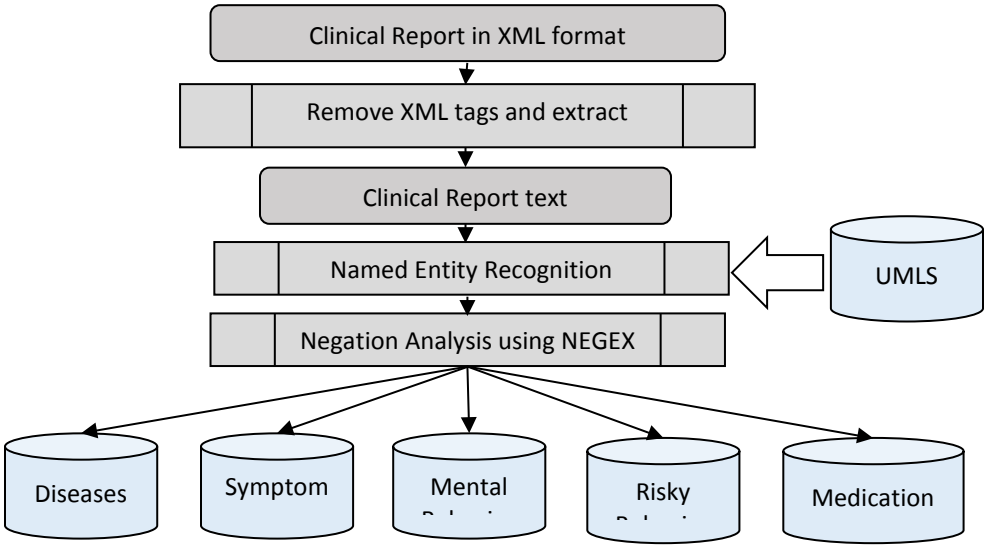


Figure 4: NLP steps applied on Clinical Reports

Once the data tables are constructed, we perform deeper analysis to compute the comorbid conditions of the diseases. For this, we use the *tf-idf* (term frequency – inverse document frequency) vector space model [9] to identify the significantly co-occurring diseases. The *tf-idf*

model is considered to be an effective text mining model that provides the importance of a term/word to a document in a collection of documents. This model uses the concept of relevance and co-occurrence of terms. Equation (1) gives the relevance of a term *j* w.r.t. a document *i*,

$$w_{ij} = t_{ij} * \lg(\frac{N}{N_j})$$

Equation 1

where w_{ij} = relevance of term *j* in the patient record *i*; t_{ij} = term frequency of term *j* in in the patient record *i*; N_j = frequency of records for term *j*; *N* = total number of records (*N*=325791). A particular term is more relevant *w.r.t.* a record if it appears more frequently in the record and appears in fewer numbers of records in the total records set. An association weight/score is attached with every association between a pair of terms [10]. This is given by A_j

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg(\frac{N}{N_j}) * t_{ik} * \lg(\frac{N}{N_k})$$

Equation 2

This is essentially a product of the relevance of each of the pair of terms over the entire records set *N*. The association score is 0 if the terms do not co-occur in any of the *N* records. Associations with non-zero scores are considered to be associated to the term.
After applying basic level processing on the reports, the clinical content from the reports was subjected to NER. UMLS was used for NER to identify the diseases, symptoms, mental behavior, risky behavior and medication terms from the 325791 reports. The total number of terms extracted for each category is given in Table 2. Figure 5 shows the most commonly occurring diseases with the number of reports in which they were found.

Term Type	Number of terms extracted using NLP
Diseases	7988
Symptoms	10803
Mental Behavior	712
Risky Behavior	244
Medications	5721

Table 2: Total terms identified by NLP

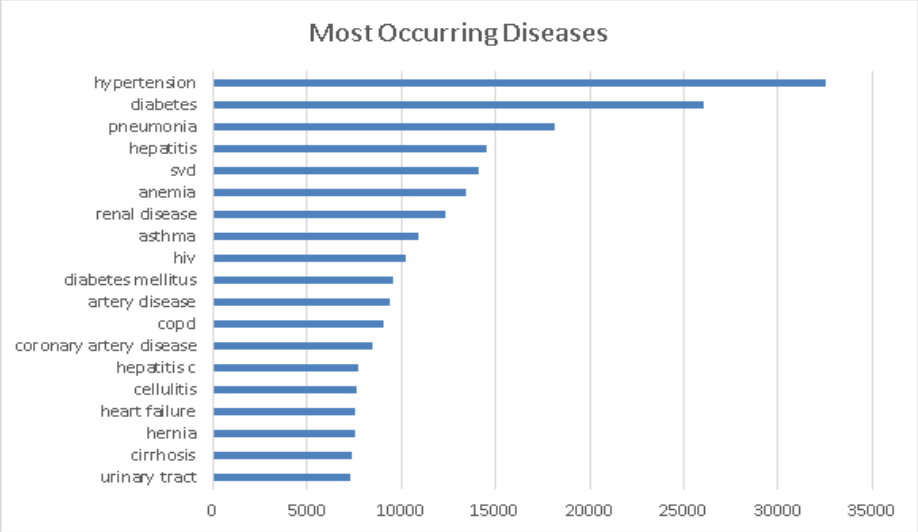


Figure 5: Most commonly occurring diseases and corresponding number of reports

The top 10 diseases were analyzed using the *tf-idf* model to identify comorbidity of the diseases across the 325791 reports. To achieve this, we compute the pair-wise significance of each disease with all the corresponding conditions, (i.e., the symptoms, mental behavior, risky behavior and medications) using Equation 2. Table 3 shows the top 10 diseases and the corresponding conditions.

We also analyzed the most common conditions that occurred with these diseases (Table 4). It was interesting to find that well known behaviors such as “smoking”, “depression” and “tobacco use” were amongst the commonly occurring conditions.

Disease Name	Diseases	Symptoms	Mental Behavior	Risky Behavior	Medications
hypertension	diabetes, renal disease, pulmonary hypertension, artery disease,	chest pain, nausea, vomiting, dyspnea, abdominal pain, weakness,	abuse, depression, dementia, anxiety, altered mental status, drug use,	smoking, tobacco use, compliance, impression, drinking, lying,	insulin, hepatitis, tobacco, oxygen, glucose, lasix,
diabetes	diabetes mellitus, hypertension, artery disease, renal disease,	nausea, vomiting, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, drug use,	smoking, compliance, tobacco use, impression, drinking,	insulin, glucose, tobacco, hepatitis, humulin,
pneumonia	lower lobe pneumonia, aspiration pneumonia, copd,	shortness of breath, chest pain, dyspnea, chills, vomiting,	abuse, dementia, aggressive, confusion,	smoking, impression, drinking, tobacco use, compliance,	oxygen, avelox, albuterol, prednisone, levaquin,
hepatitis	hepatitis c, hepatitis b, cirrhosis, liver disease, encephalopathy,	nausea, abdominal pain, vomiting, diarrhea, chills,	abuse, dependence, confusion, drug use, opiate, depression,	smoking, drinking, tobacco use, illicit drug use,	hepatitis, hepatitis b, prograf, lactulose, ammonia, antibody,
svd	gbs, pcc, ofc, strep, hep, external genitalia,	prn pain, constipation, cramping, headache,	abuse, drug use, depression, substance, substance abuse,	smokes, illicit drug use, smoking, tobacco use,	micronor, vitamin, antibody, ibuprofen, stool softener,
anemia	renal failure, diabetes, hypertension, renal disease, hepatitis, heart failure,	nausea, abdominal pain, vomiting, chest pain, fatigue, weakness,	abuse, depression, anxiety, dementia, confusion, altered mental status,	smoking, drinking, impression, tobacco use, compliance,	iron, vitamin, hepatitis, coumadin, oxygen, prednisone,
renal disease	end-stage renal disease, end stage renal disease, diabetes, hypertension, artery disease,	nausea, vomiting, chest pain, abdominal pain, chills, shortness of breath,	altered mental status, abuse, confusion, dementia, depression, confused,	smoking, compliance, impression, tobacco use, illicit drug use, drinking,	calcium, insulin, glucose, coumadin, hepatitis, bicarbonate,
asthma	pneumonia, diabetes, copd, hypertension, airway disease,	wheezing, shortness of breath, wheezes, coughing, dyspnea,	abuse, depression, mdi, anxiety, drug use, aggressive,	smoking, impression, drinking, tobacco use, crying,	albuterol, prednisone, medrol, oxygen, atrovent, advair,
hiv	aids, pneumonia, hepatitis, infectious disease, herpes, meningitis,	nausea, vomiting, diarrhea, abdominal pain, headache, weakness,	abuse, depression, schizophrenia, drug use, dementia, dependence,	compliance, smoking, drinking, impression, lying, tobacco use	hepatitis, bactrim, vitamin, cocaine, acetaminophen, hepatitis b,
diabetes mellitus	diabetes, hypertension, artery disease, renal disease,	vomiting, nausea, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, dementia,	smoking, tobacco use, compliance, illicit drug use,	insulin, glucose, humulin, tobacco, hepatitis,

Table 3: Comorbid conditions with top 10 most occurring diseases

Symptom	Occurance	Mental Behavior	Occurance	Risky Behavior	Occurance
vomiting	11	abuse	11	smoking	11
abdominal pain	10	depression	11	tobacco use	10
chest pain	10	anxiety	10	compliance	10
nausea	10	drug use	10	impression	10
weakness	9	altered mental status	9	drinking	9
diarrhea	8	confusion	9	illicit drug use	6
dyspnea	8	dementia	8	lying	6
shortness of breath	8	confused	5	crying	2
chills	7	drug abuse	5	grunting	1
headache	4	aggressive	4	marijuana	1
constipation	2	dependence	3	sobriety	1

Table 4: Most comorbid behaviors for the top 10 diseases

b. Association mining

In layer 3, we compute and store two types of associations. The first type is the conditional probability, or rule confidence, between two entities. Given two different entities i and j, the rule confidence between *i* and *j* is computed as

$$\text{Rule\_confidence}(i, j) = \frac{|i \wedge j|}{|i|}$$

in which $|i \wedge j|$ is the number of patients showing both entities *i* and *j* and | *i* | is the number of patients showing entity *i*. The second type of association shows the happen-before relationship between entities *i* and *j*, and is computed as the probability that entity *i* detection time is before entity *j* detection time

$$\text{Happen\_before}(i, j) = \frac{|i \text{ before } j|}{|i \wedge j|}$$

in which $|i \text{ before } j|$ is the number of showing *i* before showing *j*. We only compute localized association when then number of patients in the location is above 1000.

The database contains significant associations which are not widely reported in literature, such as Antidarrheal treatment and runny nose symptom (confidence: 0.73), screla and Tylenol treatment (confidence 0.70), posturing and Mortin treatment (confidence: 0.81), etc. Table 5 shows part of the association rules in tabular format. The premise and conclusion of the rule is shown in the table, with the quality measures of each rule including the support, confidence, Laplace, Gain, p-s, lift and Conviction. We are working with domain experts on evaluating the association rules and tuning the parameters to produce optimum result.

No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Convic...
5	SYPHILIS	HUMAN IMMUNODEFICIENCY VIRUS	0.010	0.286	0.976	-0.060	0.007	3.468	1.285
6	HEPATITIS A	HEPATITIS C	0.028	0.294	0.939	-0.162	0.017	2.626	1.258
7	HEPATITIS A	HEPATITIS B	0.032	0.332	0.942	-0.159	0.009	1.423	1.148
8	MUMPS	CHICKENPOX	0.010	0.348	0.981	-0.050	0.008	4.377	1.413
9	MEASLES	CHICKENPOX	0.033	0.368	0.948	-0.145	0.026	4.628	1.457
10	CHICKENPOX	HEPATITIS B	0.029	0.370	0.954	-0.130	0.011	1.589	1.218
11	MEASLES	HEPATITIS B	0.036	0.403	0.951	-0.142	0.015	1.726	1.284
12	HEPATITIS C	HEPATITIS B	0.045	0.404	0.940	-0.179	0.019	1.732	1.287
13	CHICKENPOX	MEASLES	0.033	0.412	0.957	-0.126	0.026	4.628	1.548
14	ENTEROCOCCUS VANCOMYCIN-RESISTANT	STAPHYLOCOCCUS METHICILLIN-RESISTANT	0.019	0.442	0.977	-0.067	0.014	3.847	1.586
15	MYCOBACTERIUM NON-TB	AFB UNDETERMINED	0.011	0.450	0.987	-0.038	0.011	31.475	1.793
16	TRICHOMONIASIS	CHLAMYDIA INFECTION	0.020	0.493	0.981	-0.060	0.010	2.098	1.509
17	MUMPS	HEPATITIS B	0.015	0.497	0.985	-0.045	0.008	2.129	1.523
18	MUMPS	MEASLES	0.016	0.539	0.987	-0.044	0.014	6.065	1.978
19	CHLAMYDIA INFECTION	GONORRHEA	0.142	0.604	0.925	-0.328	0.094	2.932	2.007
20	GONORRHEA	CHLAMYDIA INFECTION	0.142	0.689	0.947	-0.270	0.094	2.932	2.460
21	AFB UNDETERMINED	MYCOBACTERIUM NON-TB	0.011	0.764	0.997	-0.018	0.011	31.475	4.143
22	TRACHOMA	CHLAMYDIA INFECTION	0.014	0.881	0.998	-0.018	0.010	3.749	6.436

Table 5: Association Rules among Diseases

c. Clustering analysis

We developed a co-clustering algorithm to cluster both diseases and text-mining terms to discover potential combinations of both diseases and terminologies, which could be disease subtypes or imply new biomedical patterns. The algorithm iteratively and partially [3] allocates the diseases or terms into clusters based on the rule confidence attributes. Let *K* be the number of clusters. To reallocate the diseases given the clustering allocation of terms, we select the *k* giving the maximum affinity score (*as* score) of disease *i* on cluster *k*. The *as* score is computed as

$$as(i,k) = \sum_{\forall j \in C_k} \left(p(i \mid j) - \overline{p(l \mid j)}_{l} \right)$$

in which C_k denotes the cluster *k*, *j* is the index of the term and $\overline{p(l \mid j)}_{l}$ is the mean of the associations given term *j*. $\overline{p(l \mid j)}_{l}$ is the repulse factor to prevent the case when all diseases and terms falls in one cluster. The process to reallocate the terms is similar to the diseases allocation process.

The algorithm can be executed in parallel by using a master-assistant computational model to improve efficiency. When reallocating diseases, the master routine sends the term-cluster allocations to all assistants and assigns the disease subsets for each assistant to reallocates. The assistants send the disease allocation results for the master routine for later use in terms allocation. We terminate the iterative allocation steps until the number of diseases/terms adopting new cluster is small and the clusters become stable.

We found several cluster containing close relationships between diseases and terminologies, such as {Biliary Sludge, HFA, Macrocytosis, Paroxysmal, Pseudogout, back discomfort, betimol, hesitancy, Intron A}, {Gastric Polyps, Kidney failure, antral, benefix, benicar}, {Duodenal Ulcer, Helicobacter Pylori, Malabsorpition, amylase, antimetics} and {appetite lost, immunoglobulin, retrovir} , etc. Some clusters highly correspond to specific diseases or medical processes. For example, appetite lost, immunoglobulin, retrovir are HIV related symptoms. Meanwhile, some clusters contain diseases and terms associated with several medical processes. For example, we found a cluster including gastrointestinal terms (Duodenal ulcer, Helicobacter pylori, Malabsorption, acyclovir, amylase and antiemetics), additive behavior (drinking, lortab andmarijuana) and cancer (methotrexate, vincristine and zofran). This cluster may suggest negative impact of addictive behavior toward digestive system. The appearance of cancer drugs

in this cluster could raise a research question about the impact of additive behavior toward the metabolism process, which will further affect the cancer drug efficiency.

We construct the sequence of disease/term occurrence based on rule_confidence and happen_before association. Only associations with rule_confidence and happen_before association greater than certain threshold and covering at least 50 patients are included to construct the sequence and visualization. Due to the limited number of text records showing the test date, we only applied sequential pattern mining on disease association.
We found 105 disease-associations satisfying all 3 criteria about rule_confidence, begin_before_end and coverage to construct the frequent sequential disease patterns. We found 3 groups of sequences in the NCD data. The first group contains only one sequence Hyperplenism, Annemia. The second group contains 5 diseases: Fibrosis Pulmonary, Staphylococcus Methicillin-resistant, Biliary Stricture, Cycsticercosis and Meconium Ileus, in which Fibrosis pulmonary and Meconium ileus stay at the triggering position. This disease group may raise additional research questions since these diseases occur at different organs. The last group is marked by Hepatitis A and 56 other diseases staying at the triggering position of Hepatitis A.

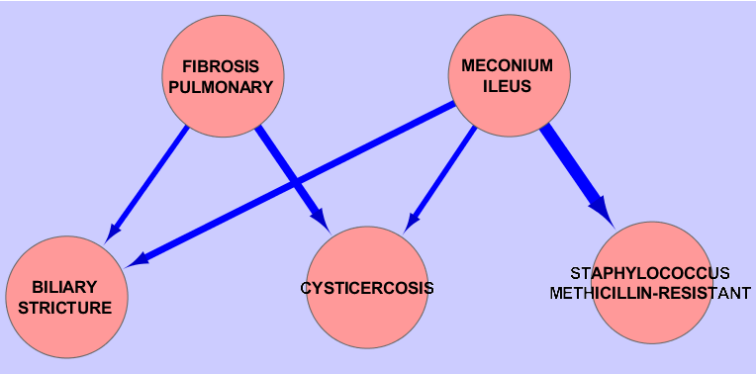


Figure 6. Fibrosis/Meconium-ileus sequence

We have developed a suite of visualization algorithms for the interactive visual exploration of the health data represented in the concept space database.

The opening visualization (Fig 7) is an associative graph of the diseases and other terms from association mining. A spring-embedder algorithm is used to layout the graph nodes:

$$E_s = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} k(d(i,j) - s(i,j))^2$$

Where *d(i,j)* is the 2D Euclidean distance of two nodes, and *s(i,j)* is a similarity metric of two nodes representing the heuristic of the layout. Edge thickness indicates the strength of association, and node size can reflects the number of other nodes to which a given node has a significant association, or the total occurrence of a term (e.g. disease) in the dataset. Nodes can be selected, and the graph will be quickly redrawn (Fig 8) to only show other nodes which have significant association to the selected nodes. The association graph is also used as a platform for the users to select terms for additional visualizations.

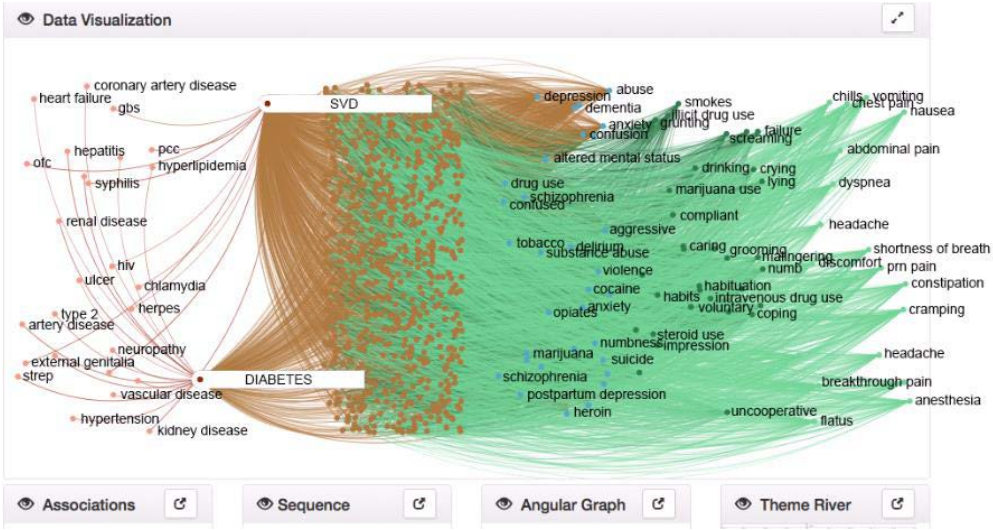


Figure 7: Associative Graph A

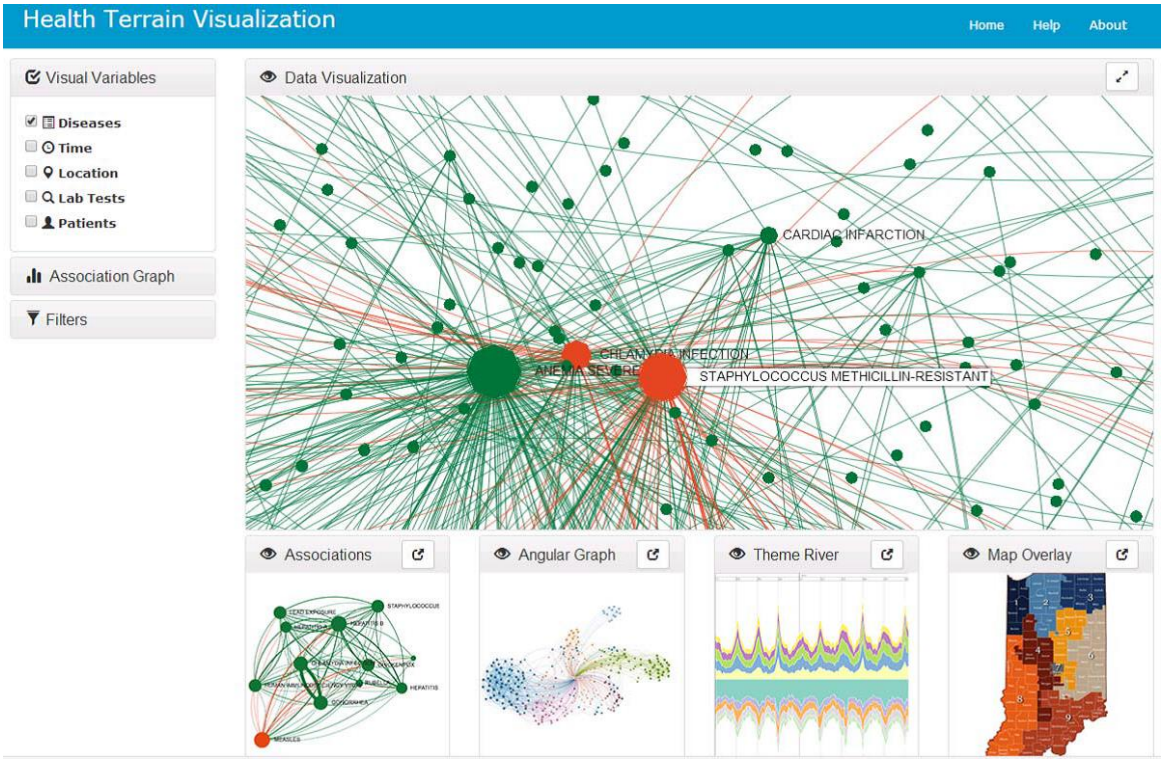


Figure 8: Associative Graph B

A visualization (Fig. 9) based on the classic theme river [11] allows the side-by-side comparison of streams of disease occurrences over time. Sudden or interesting increases or decreases are easily spotted in this interactive graphic.

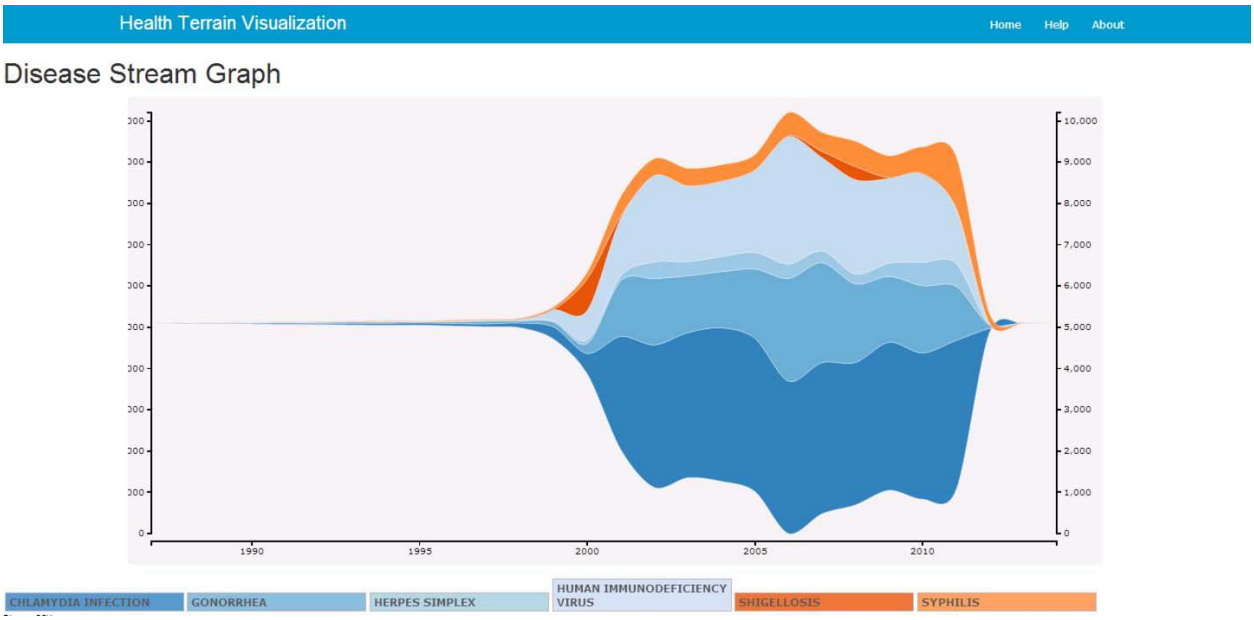


Figure 9: Classic Theme River Visualization

The patient graph (Fig. 10) is a new visualization method that allows the users to see the total patient occurrences for a selected time sequence and any number of diseases. Each node represents an occurrence of a disease for a given patient, and lines across the graph show multiple occurrences for the same patient. We will change these lines to curves to better represent the flow of time axis. Alternating bands of gray and white represent different diseases, and time increases as you move along the arc of the concentric circles. Other variables such as sex and race are indicated by the color and shape of the occurrence nodes.

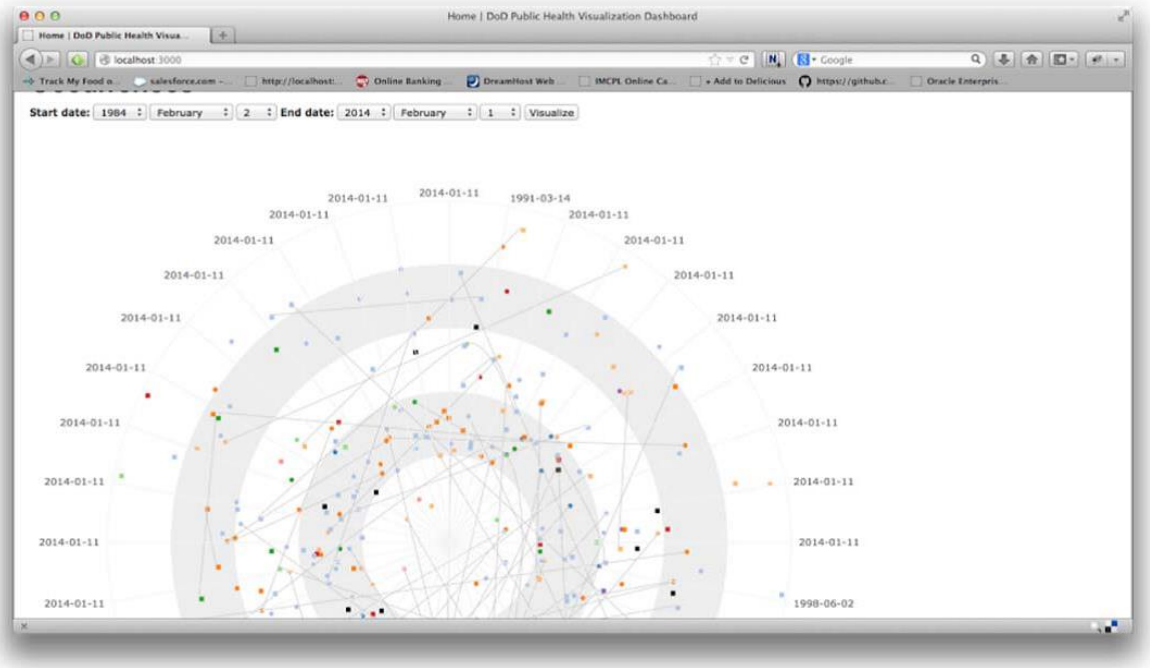


Figure 10: Patient Graph

A classic heat map (Fig. 11) gives users a view into disease (or other terms) distributions across a geographic area. Interactivity based on selected diseases, time ranges and zooming capabilities provides interesting drill-down details.

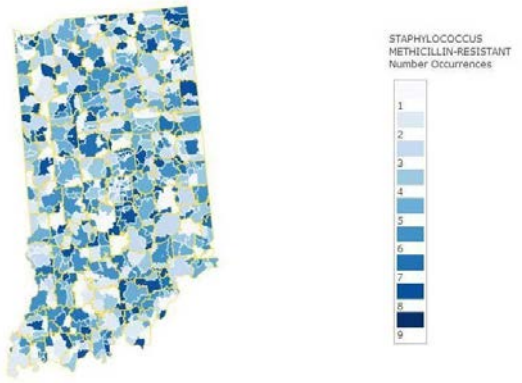


Figure 11: Heat map

A terrain surface method will be applied to visualize health term distributions over the Indiana map. One or more focal concept terms (e.g. a disease) can be defined as response variables, and represented as height and other visual parameters, such as color. A smooth surface will then be interpolated over the attributes at different locations within the state of Indiana (typically zip codes). The Shepard interpolation method will be used,

$$d = \sum_{i=0}^{n-1} (1/r_i)^2 \cdot d_i \Big/ \sum_{i=0}^{n-1} (1/r_i)^2$$

where *d* is the height of an arbitrary point *P* within the Indiana map, *di* are the known heights (attributes) at the centers, *Ci*, of the regions (e.g. zip codes or counties), and *ri* are the distances between *P* and *Ci*. A 2D image of the state map will be used to restrict the surface to be within the Indiana border. Once the surface is constructed, horizontal cross sectional contours can be generated to identify the geographical regions that response more sensitively to the given term. This technique will be implemented as a variation of our previous work on GeneTerrain [12]. The main challenge in using terrain surface for health data is how to visually represent multiple attributes (e.g. the associated diseases or symptoms of a given disease) in one terrain surface. A similar problem is how to represent the time-series data (multiple time intervals) in one terrain surface. We have developed a novel spatiotemporal terrain surface algorithm which employs offset contours to depict multiple attributes or a time-series attributes in each geographical region, as shown in the Figure below. This algorithm is still under development, and will be integrated within the GUI soon.

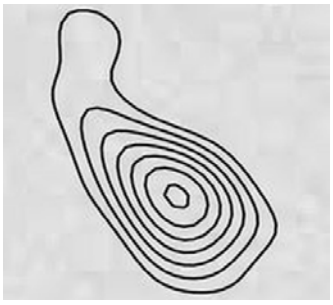


Figure 12: Spatiotemporal Terrain Surface

3. System Design and Implementation

We designed and implemented the initial structure and framework of HelthTerrain visualization. Initial plans had been to develop an installed executable application written in C++ and utilizing OpenGL for interactive visualizations. However, after some initial research and experimentation in the capabilities of modern web browsers (Google Chrome, Mozilla Firefox and Apple Safari) for 2D and 3D graphics, we came to the conclusion that WebGL in an HTML5 canvas would provide sufficient technical and graphical capabilities we need while appealing to a much broader potential user base with an established and maturing set of user experience patterns. Once focused on the web, we settled on an architecture pattern based primarily on the Ruby on Rails (RoR) framework for delivering web applications with AJAX services and a classic Model-View-Controller architecture. Ruby and Rails were picked as our server-side language and framework of choice for their elegant syntax, vibrant open source community, and ease of use. The application itself is 3-part:

a. A MySQL relational database containing the results of offline text mining and statistical analysis research on the health data set provided to us by our partners at Regenstrief Institute.
b. A server-side RoR application for querying, modeling and manipulating data in the relational database.
c. An HTML/CSS/JavaScript web GUI.

The user interface is a modern web GUI utilizing a combination of form submission and RESTful service calls to query and retrieve data in various data delivery formats such as Extensible Markup Language (XML) and JavaScript Object Notation (JSON). Interactivity is a primary goal as we seek to both visualize our data and provide opportunities for novel visual exploration and analysis.

The visualizations themselves utilize HTML, CSS, SVG, and WebGL technologies with a number of open-source JavaScript libraries such as sigma.js, d3.js, jquery.js and three.js for drawing, displaying and interacting with the data and graphics.

Figure 13: Visualization Interface

Key Research Accomplishments

- Designed and implemented a MySQL relational database, as a representation of the concept space, which is derived from the NCD dataset by data mining and text mining algorithms.
- Natural Language Processing techniques were carried out to process 325791 clinical notes to extract new terms including diseases, symptoms, and mental and risky behaviors.
- Data mining techniques were applied to extract associations between terms in the concept space, and to discover new cluster terms.
- Designed and partially implemented a suite of interactive visualization algorithms that allows the users to interactively explore the data based on the user selected terms and filters.
- Designed and implemented a web based graphical user interface for the prototype system, and successfully integrated the programming interfaces between the user interface, visualization, and the database.

Reportable Outcomes

Manuscripts, Abstracts, Presentation (a copy of each is included in the appendices)

T. Nguyen, A. Krishnan, S. Bloomquist, J. Keiper, W. Li, S. Grannis, Y. Xia, S. Fang, M. Palakal. A Knowledge Discovery System for Notifiable Condition Detector Data. Submitted to AMIA 2014.(Submitted)

J. Keiper, Y. Xia, S. Fang, M. Palakal, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper, S. Grannis. Use Cases for Public Health Data Visualization. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC. Oct. 2013.

Y. Xia, S. Fang, M. Palakal, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper, S. Grannis. Data Exploration of a Notifiable Condition Detector System. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC, Oct. 2013.

M. Palakal, S. Fang, Y. Xia, S. Grannis, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper. Detecting Comorbidity of Chlamydia from Clinical Reports. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC. Oct. 2013.

Informatics such as databases

We have constructed 5 ontologies, one for each category - Disease, Symptom, Mental behavior, Risky Behavior and Medication based on the data extracted from the clinical notes. This helps us in eliminating noise and providing structure to our findings.

We have built a specialized database for the storage and real time query of the concept space and the NCD dataset. The database is general enough to be adapted to any other health care data and extensions of the concept space.

Conclusion

We have made significant overall progress in this project, including: (1) creating a concept space definition, which represents a schema tailored to support diverse visualizations and provides a uniform ontology that allows the system to be leveraged for many types of health care datasets through individually designed text and data mining procedures; (2) designing and implementing data and text mining analytics and visualization algorithms; and (3) developing a flexible prototype system for using our analytics and visualization framework to explore large-scale, real-world health data. These three components are integrated in a generalizable browser-based graphical interface, which enables flexible and free-form data exploration and hypothesis discovery and also a more flexible distribution of the resulting software. We have completed the majority of algorithm development and implementation; implementation and testing of a few remaining advanced visualization techniques are outstanding. The system has received favorable initial feedback from users, and we believe it has potential as an open source tool to support health data visualization tasks. In the second year of the project, we will focus efforts on more formally engaging clinical and public health stakeholders to test and iteratively refine the visual analytics use case.

References

Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/

Fighting disease outbreaks with two-way health information exchange, last retrieved from http://newsinfo.iu.edu/news/page/normal/11948.html

Gardner RM, Golubjatnikov OK, Laub RM, Jacobson JT, Evans RS. Computer-critiqued blood ordering using the HELP system. Comput Biomed Res 1990;23:514-28.

J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury and M. Roland. *Defining Comorbidity: Implications for Understanding Health and Health Services*. Ann Fam Med. 2009 July; 7(4): 357–363. Pubmed ID: 19597174

L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng G and W. A. Kibbe. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012 Jan. Pubmed ID: 22080554

P. Shannon, A. Markiel, O. Ozier, N.S Baliga, J.T Wang, E. Ramage, N. Amin, B. Schwikowski, T. Ideker. *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. Genome Res. 2003 Nov;13(11):2498-504. Pubmed ID: 14597658

B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–11

Chapman , W.; Bridewell , ; Hanbury , ; Cooper , G. F.; Buchanan , G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics* 2001, *34* (5), 301–310.

Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/

Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342

S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *Visualization and Computer Graphics, IEEE Transactions*, vol. 8, pp. 9-20, 2002

You, Q., Fang, S., Chen, J. GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks. Journal of Information Visualization, 2010; 9:1, 1-12.

Appendices

A Knowledge Discovery System for Notifiable Condition Detector Data

Thanh Minh Nguyen[1] , Anand Krishnan[2], Sam Bloomquist[1], Jeremy Keiper[3], Weizhi Li[2], Shaun Grannis, MD[3], Yuni Xia, PhD[1], Shiaofen Fang, PhD[1], Mathew Palakal, PhD[2], [1]Department of Computer Science, Indiana University – Purdue University, Indianapolis; [2]School of Informatics and Computing, Indiana University – Purdue University, Indianapolis; [3]Regenstrief Institute, Indianapolis, IN

Abstract

The Notifiable Condition Detector (NCD) system is an automated electronic lab reporting (ELR) and case-notification system developed by Regenstrief Institute. It has been used in Indiana for over ten years to report laboratory results for the detection of notifiable conditions such as novel H1N1 influenza, sexually transmitted diseases, lead poisoning, and salmonella[1]. In this paper, we present a knowledge discovery system which integrates database, text mining and data mining on the NCD data. We show that the knowledge discovery system could not only discover new associations among terminologies in health science but also enhance friendly visualization application to show more significant and useful data to users.

Introduction

The Regenstrief Institute implemented and maintains an HIE-based, automated electronic lab reporting (ELR) and case-notification system for over ten years in Indiana State Marion County. The Notifiable Condition Detector (NCD) System uses a standards-based messaging and vocabulary infrastructure that includes Health Level Seven (HL7) and Logical Observation Identifiers Names and Codes (LOINC). The NCD receives real-time HL7 version 2 clinical transactions daily, including diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations[2].

The NCD, now operational in Indiana, automatically detects positive cases of indicated conditions and forwards alerts to local and state health departments for review and possible follow up. These alerts assist public health agencies to perform population health monitoring more efficiently and effectively.

Data Summary

The dataset we received from Regenstrief NCD system contains 833,710 observations. The dataset has been de-identified, with patient age and zip code pseudonymized. The dataset contains 47 columns including patient pseudo ID, condition name, test result name, test result value, test normal range, patient race, patient gender, etc. The missing data rate for columns varied substantially from 0% for column Patient Pseudo_ID to over 70% for column Test_Abnormal_Flag.

Methods

The NCD receives clinical data that includes the diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations. But much of the information pertaining the patients' condition is available in the clinical reports. Mining these reports can provide a bigger picture of various other conditions that the patient experienced during his/her treatment. This information can provide valuable insights on the patients' socioeconomic condition, behavior risk factors, environmental factors and genetic information (family history). Natural Language Processing (NLP) provides a means to augment the NCD data analytics with the information discovered from these clinical reports.

Text Mining Process

NLP techniques were carried out to process 325791 clinical notes that contain patient discharge summaries, laboratory reports, patient history, etc. Although these records are de-identified due to which the patient specific information are absent, a pseudo-patient Id has been provided to help process the reports. Basic processing of the reports was performed for converting the clinical notes from XML format to simple text format and sentence splitting. Advanced level NLP was applied in the form of named entity recognition (NER) for extracting diseases, symptoms, mental behavior, risky behavior and medication information from the reports. This was done with the help of UMLS[3] database which is a repository of clinical and health related terms. Once the entities were extracted using NER, negation analysis was applied using NEGEX algorithm[4] to remove negated terms. Figure 1 shows the process that was used in extracting this vital information from the reports.

The advantage of using UMLS is that all variations of clinical terms get captured that provide a large set of terms available for further analysis. For example, clinical notes that indicate "Hepatitis" contains terms like "Hepatitis", "Hepatitis B", "Hep", "Hep B" etc. The large number of terms extracted contains different occurrences of the same diseases, symptoms, etc. We apply stemming and grouping algorithms to group these terms to reduce the total number of terms. The identified terms are stored in different data tables and joined using the pseudo-patient Id.

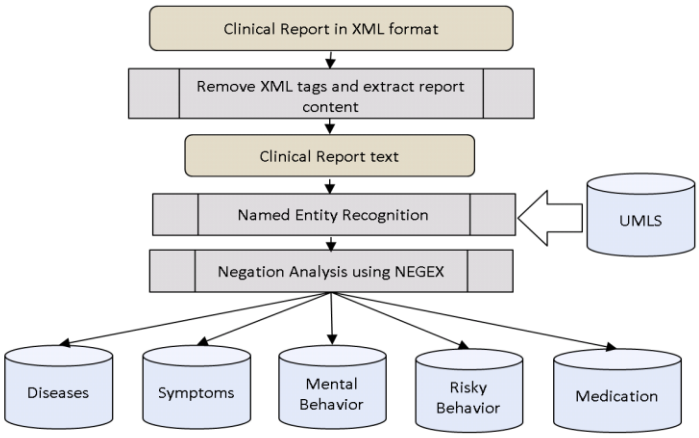


Figure 1. NLP steps applied on clinical reports

Comorbidity Analysis

Once the data tables are constructed, we perform deeper analysis to compute the comorbid conditions of the diseases. For this, we use the *tf-idf* (term frequency – inverse document frequency) vector space model[5] to identify the significantly co-occurring diseases. The *tf-idf* model is considered to be an effective text mining model that provides the importance of a term/word to a document in a collection of documents. This model uses the concept of relevance and co-occurrence of terms. Equation (1) gives the relevance of a term *j* w.r.t. a document *i*,

$$w_{ij} = t_{ij} * \lg(\frac{N}{N_j}) \quad (1)$$

where w_{ij} = relevance of term *j* in the patient record *i*; t_{ij} = term frequency of term *j* in in the patient record *i*; N_j = frequency of records for term *j*; *N* = total number of records (*N*=325791).

A particular term is more relevant *w.r.t.* a record if it appears more frequently in the record and appears in fewer numbers of records in the total records set. An association weight/score is attached with every association between a pair of terms[6]. This is given by A_{jk}

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg(\frac{N}{N_j}) * t_{ik} * \lg(\frac{N}{N_k}) \quad (2)$$

This is essentially a product of the relevance of each of the pair of terms over the entire records set *N*. The association score is 0 if the terms do not co-occur in any of the *N* records. Associations with non-zero scores are considered to be associated to the term.

Database design

Considering the visualization specific requirements, we designed a 3-layer database model (Figure 2) to store the NCD and supported text dataset. The first layer contains 4 base tables for 4 entities: patient, disease, term and location. The term table has 4 subcategories: mental behavior, risky behavior, medication and symptom. The second layer contains the associations between the patient entity and other entities. Therefore, the NCD and supported text dataset could be recovered by joining tables in the second layer over the patient table. The third layer contains indirect associations between disease, term and location. Constructing the third layer require data mining techniques. We decide the specific tables and types of associations based on frequent queries requested from the visualization application. In addition, there are tables to store query results such as counting the number of patients having disease x at location y to avoid database scan during visualization execution. By caching the results of queries, the database

can avoid having to repeat the potentially time-consuming and intensive operations (for example, sorting/aggregation, joins etc) that generated the query results and speed up rendering and visualization. The association could be both at global scale and at local scale.

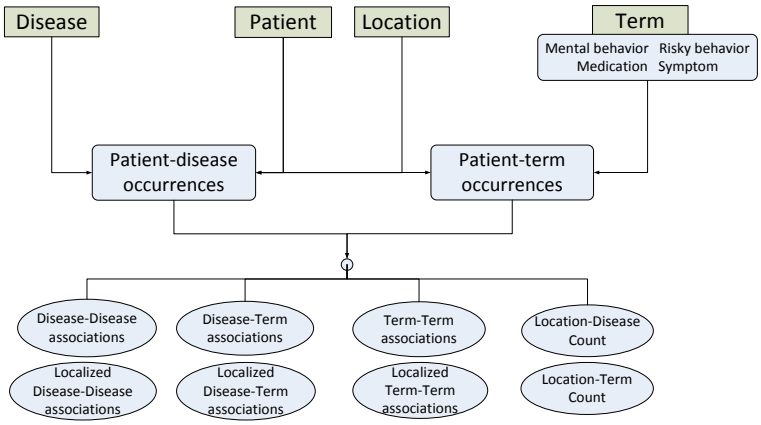


Figure 2. Database model

Data cleansing

There are cases in which one PseudoID may match to more than one patient. To avoid this error, we create three checkpoints to ensure that patients sharing the same PseudoID are indeed the same patient. Three checkpoints are gender, race, and date of birth. For gender checkpoints, if two patients show different genders in the records and none of the genders are NULL or U (unknown), then the gender checkpoint fails. The race checkpoint works in a similarly way. For the date of birth checkpoint, we convert the date of birth into yyyy-mm-dd format and perform longest common subsequence matching[7]. If the ratio between the length of the longest common subsequence over the length of yyyy-mm-dd format is less than a certain threshold, we will decide that the date of birth checkpoint fail. Different entries are from the same patient if and only if all three checkpoints pass.

In addition, we discovered cases when the terms mined from the text having different canonical name but representing the same entity, such as 'pain', 'pains' and 'x_pains'. We applied the procedure similar to patient data cleansing to eliminate duplications among these terms.

Association Mining

In layer 3, we compute and store two types of associations. The first type is the conditional probability, or rule confidence, between two entities. Given two different entities i and j, the rule confidence between *i* and *j* is computed as

$$\text{Rule\_confidence}(i, j) = \frac{\left|i \wedge j\right|}{|i|} \quad (9)$$

in which $\left|i \wedge j\right|$ is the number of patients showing both entities *i* and *j* and $|i|$ is the number of patients showing entity *i*. The second type of association shows the happen-before relationship between entities *i* and *j*, and is computed as the probability that entity *i* detection time is before entity *j* detection time

$$\text{Happen\_before}(i, j) = \frac{\left|i \text{ before } j\right|}{|i \wedge j|}$$

in which $\left|i \text{ before } j\right|$ is the number of showing *i* before showing *j*. We only compute localized association when then number of patients in the location is above a minimal threshold. The reason of enforcing the minimal patient threshold is to ensure the statistical significance of the associations identified.

Clustering Analysis

We develop a co-clustering algorithm to cluster both diseases and text-mining terms to discover potential combinations of both diseases and terminologies, which could be disease subtypes or imply new biomedical patterns. The algorithm iteratively and partially[10] allocates the diseases or terms into clusters based on the rule confidence attributes. Let *K* be the number of clusters. To reallocate the diseases given the clustering allocation of terms, we select the *k* giving the maximum affinity score (*as* score) of disease *i* on cluster *k*. The *as* score is computed as

$$as(i,k) = \sum_{\forall j \in C_k} \left(p(i \mid j) - \overline{p(l \mid j)}_{l} \right)$$

in which C_k denotes the cluster *k*, *j* is the index of the term and $\overline{p(l \mid j)}_{l}$ is the mean of the associations given term *j*. $\overline{p(l \mid j)}_{l}$ is the repulse factor to prevent the case when all diseases and terms falls in one cluster. The process to reallocate the terms is similar to the diseases allocation process.

The algorithm can be executed in parallel by using a master-assistant computational model and. When reallocating diseases, the master routine sends the term-cluster allocations to all assistants and assigns the disease subsets for each assistant to reallocates. The assistants send the disease allocation results for the master routine for later use in terms allocation. We terminate the iterative allocation steps until the number of diseases/terms adopting new cluster is small and the clusters become stable.

Sequential Pattern Mining

We construct the sequence of disease/term occurrence based on rule_confidence and happen_before association. Only associations with rule_confidence and happen_before association greater than certain threshold and covering at least 50 patients are included to construct the sequence and visualization. Due to the limited number of text records showing the test date, we only applied sequential pattern mining on disease association.

Results

Text Mining Results

After applying basic level processing on the reports, the clinical content from the reports was subjected to NER. UMLS was used for NER to identify the diseases, symptoms, mental behavior, risky behavior and medication terms from the 325791 reports. The total number of terms extracted for each category is given in Table 1. Figure 3 shows the most commonly occurring diseases with the number of reports in which they were found.

The top 10 diseases were analyzed using the tf-idf model to identify comorbidity of the diseases across the 325791 reports. To achieve this, we compute the pair-wise significance of each disease with all the corresponding conditions, (i.e., the symptoms, mental behavior, risky behavior and medications) using Equation 2. Table 2 shows the top 10 diseases and the corresponding conditions.

Table 1: Total terms identified by NLP

Term Type	Number of terms extracted using NLP
Diseases	7988
Symptoms	10803
Mental Behavior	712
Risky Behavior	244
Medications	5721

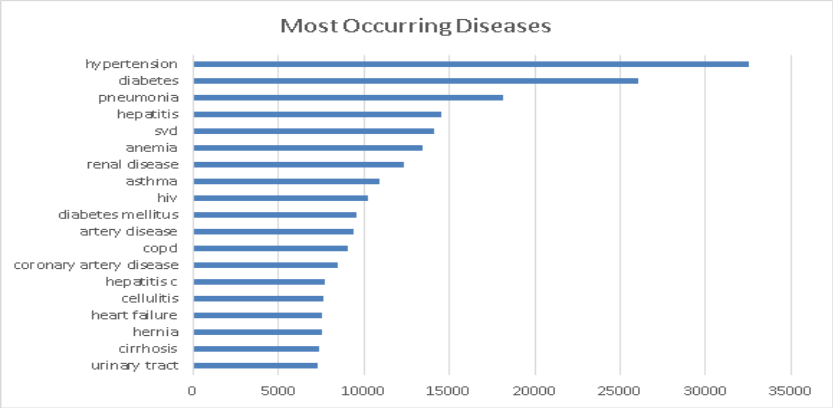


Figure 3. Most commonly occurring diseases and corresponding number of reports

We also analyzed the most common conditions that occurred with these diseases. It was interesting to find (Table 3) that well known behaviors such as "smoking", "depression" and "tobacco use" were amongst the commonly occurring conditions.

Table 2: Comorbid conditions with top 10 most occurring diseases

Disease Name	Diseases	Symptoms	Mental Behavior	Risky Behavior	Medications
hypertension	diabetes, renal disease, pulmonary hypertension, artery disease,	chest pain, nausea, vomiting, dyspnea, abdominal pain, weakness,	abuse, depression, dementia, anxiety, altered mental status, drug use,	smoking, tobacco use, compliance, impression, drinking, lying,	insulin, hepatitis, tobacco, oxygen, glucose, lasix,
diabetes	diabetes mellitus, hypertension, artery disease, renal disease,	nausea, vomiting, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, drug use,	smoking, compliance, tobacco use, impression, drinking,	insulin, glucose, tobacco, hepatitis, humulin,
pneumonia	lower lobe pneumonia, aspiration pneumonia, copd,	shortness of breath, chest pain, dyspnea, chills, vomiting,	abuse, dementia, aggressive, confusion,	smoking, impression, drinking, tobacco use, compliance,	oxygen, avelox, albuterol, prednisone, levaquin,
hepatitis	hepatitis c, hepatitis b, cirrhosis, liver disease, encephalopathy,	nausea, abdominal pain, vomiting, diarrhea, chills,	abuse, dependence, confusion, drug use, opiate, depression,	smoking, drinking, tobacco use, illicit drug use,	hepatitis, hepatitis b, prograf, lactulose, ammonia, antibody,
svd	gbs, pcc, ofc, strep, hep, external genitalia,	prn pain, constipation, cramping, headache,	abuse, drug use, depression, substance, substance abuse,	smokes, illicit drug use, smoking, tobacco use,	micronor, vitamin, antibody, ibuprofen, stool softener,
anemia	renal failure, diabetes, hypertension, renal disease, hepatitis, heart failure,	nausea, abdominal pain, vomiting, chest pain, fatigue, weakness,	abuse, depression, anxiety, dementia, confusion, altered mental status,	smoking, drinking, impression, tobacco use, compliance,	iron, vitamin, hepatitis, coumadin, oxygen, prednisone,
renal disease	end-stage renal disease, end stage renal disease, diabetes, hypertension, artery disease,	nausea, vomiting, chest pain, abdominal pain, chills, shortness of breath,	altered mental status, abuse, confusion, dementia, depression, confused,	smoking, compliance, impression, tobacco use, illicit drug use, drinking,	calcium, insulin, glucose, coumadin, hepatitis, bicarbonate,
asthma	pneumonia, diabetes, copd, hypertension, airway disease,	wheezing, shortness of breath, wheezes, coughing, dyspnea,	abuse, depression, mdi, anxiety, drug use, aggressive,	smoking, impression, drinking, tobacco use, crying,	albuterol, prednisone, medrol, oxygen, atrovent, advair,
hiv	aids, pneumonia, hepatitis, infectious disease, herpes, meningitis,	nausea, vomiting, diarrhea, abdominal pain, headache, weakness,	abuse, depression, schizophrenia, drug use, dementia, dependence,	compliance, smoking, drinking, impression, lying, tobacco use	hepatitis, bactrim, vitamin, cocaine, acetaminophen, hepatitis b,
diabetes mellitus	diabetes, hypertension, artery disease, renal disease,	vomiting, nausea, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, dementia,	smoking, tobacco use, compliance, illicit drug use,	insulin, glucose, humulin, tobacco, hepatitis,

Table 3: Most comorbid behaviors for the top 10 diseases

Symptom	Occurance	Mental Behavior	Occurance	Risky Behavior	Occurance
vomiting	11	abuse	11	smoking	11
abdominal pain	10	depression	11	tobacco use	10
chest pain	10	anxiety	10	compliance	10
nausea	10	drug use	10	impression	10
weakness	9	altered mental status	9	drinking	9
diarrhea	8	confusion	9	illicit drug use	6
dyspnea	8	dementia	8	lying	6
shortness of breath	8	confused	5	crying	2
chills	7	drug abuse	5	grunting	1
headache	4	aggressive	4	marijuana	1
constipation	2	dependence	3	sobriety	1

Database after cleansing

After cleansing, the database has 439,547 patients, 1976 diseases, 3756 locations and 3851 terms (711 symptoms, 93 risky behaviors, 200 mental behaviors and 2847 medications). At the second layer, the database contains 1,302,173 disease occurrences and 1,215,659 term occurrences. However, there are only 90,376 patients associating with at least one term. All of these patients have a least one disease. The number of patients having more than one disease is 114,820, which is later used for association mining. At the third layer, the database contains 577,888 global associations between two different diseases, 1,958,227 global associations between two different terms and 1,032,864 global associations between a disease and a term.

We first remove from the dataset duplicate rows, where the same patient reports the same condition multiple times. Then, we identified the most common diseases in the data. "Lead Exposure" is the condition that has the highest occurrence- it occurs in 256823 patients; however, lead poisoning is not common in practice. The reason "Lead Exposure" has the highest occurrence in the data is that under the state's reporting law, all laboratories performing blood lead tests are required to report the results of those tests. Therefore, even when the test result is in the normal range, the test was reported. It leads to a high number of records on "Lead Exposure" in the data, while most of the report has negative results. Other than "Lead Exposure", the most common diseases include 1)Staphylococcus Methicillin-Resistant Aureus (MRSA), 2)HIV, 3)Chlamydia Infection, 4)Hepatitis B, 5)Hepatitis C, 6)Gonorrhea, 7)Chickenpox 8)Measles 9)Hepatitis A 10) Enterococcus Vancomycin-Resistant (VRE) 11) Trichomoniasis 12) Syphilis. Figure 4 shows the number of occurrences of the most common diseases.

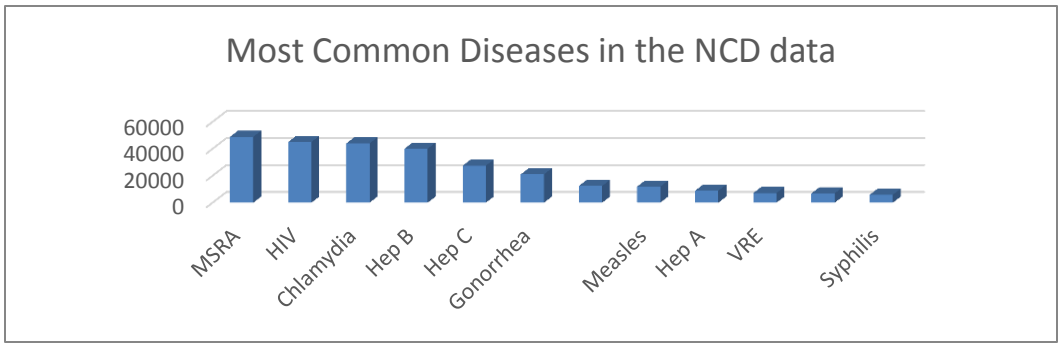


Figure 4. Most Common Diseases in the NCD dataset

Disease Distribution Across Race

We analyze the disease distribution across races. Here we compare the difference between the two largest races: white and black. The result is shown in Figure 5, with the black bar representing the occurrence percentage of each disease among black patients and the blue bar representing the occurrence percentage of disease among white patients. It shows that among black patients, CHLAMYDIA INFECTION and GONORRHEA are the most common conditions in the NCD data. TRICHOMONIASIS and SYPHILIS are also more common in black patients than in white patients. Among white patients, the most common condition is STAPHYLOCOCCUS METHICILLIN-RESISTANT AUREUS (MSRA).

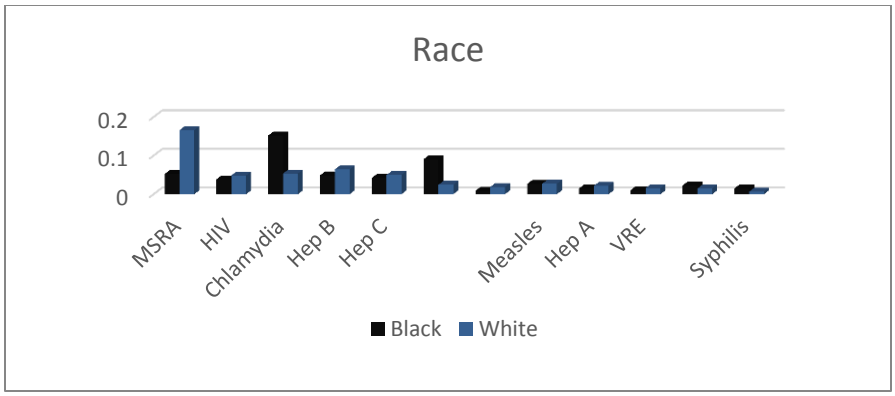


Figure 5. Diseases Distribution Across Race

Query efficiency

We test the database efficiency by three sets of common samples queries designed by visualization and health science experts. The first query set is about geographical distribution of one or a combination of diseases. The second query set retrieves strong associated diseases to a given disease. The third query set finds common diseases occurring at a given range of age. Table 4 summarizes the performance of three types of queries and suggests that the database is optimized and effectively support real time association mining.. The performance of aggregated queries is satisfactory and may be further optimized. .

Table 4: Three sets of query used for testing the database

Query set	Example	Involved tables	Runtime
1	Geographical (at city level) distribution of chlamydia	location, diseases, patient-disease occurrences	12s
2	List the diseases associating with chlamydia	diseases, associations	0.5s
3	What are the most common diseases for patient age from 20 to 40	diseases, patients, patient-disease occurrences	16s

Association Mining

The database contains significant associations which are not widely reported in literature, such as Antidarrheal treatment and runny nose symptom (confidence: 0.73), screla and Tylenol treatment (confidence 0.70), posturing and Mortin treatment (confidence: 0.81), etc.

Table 5 shows part of the association rules in tabular format. The premise and conclusion of the rule is shown in the table, with the quality measures of each rule including the support, confidence, Laplace, Gain, p-s, lift and Conviction. We are working with domain experts on evaluating the association rules and tuning the parameters to produce optimum result.

Table 5: Association Rules among Diseases

No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Convic...
5	SYPHILIS	HUMAN IMMUNODEFICIENCY VIRUS	0.010	0.286	0.976	-0.060	0.007	3.468	1.285
6	HEPATITIS A	HEPATITIS C	0.028	0.294	0.939	-0.162	0.017	2.626	1.258
7	HEPATITIS A	HEPATITIS B	0.032	0.332	0.942	-0.159	0.009	1.423	1.148
8	MUMPS	CHICKENPOX	0.010	0.348	0.981	-0.050	0.008	4.377	1.413
9	MEASLES	CHICKENPOX	0.033	0.368	0.948	-0.145	0.026	4.628	1.457
10	CHICKENPOX	HEPATITIS B	0.029	0.370	0.954	-0.130	0.011	1.589	1.218
11	MEASLES	HEPATITIS B	0.036	0.403	0.951	-0.142	0.015	1.726	1.284
12	HEPATITIS C	HEPATITIS B	0.045	0.404	0.940	-0.179	0.019	1.732	1.287
13	CHICKENPOX	MEASLES	0.033	0.412	0.957	-0.126	0.026	4.628	1.548
14	ENTEROCOCCUS VANCOMYCIN-RESISTANT	STAPHYLOCOCCUS METHICILLIN-RESISTANT	0.019	0.442	0.977	-0.067	0.014	3.847	1.586
15	MYCOBACTERIUM NON-TB	AFB UNDETERMINED	0.011	0.450	0.987	-0.038	0.011	31.475	1.793
16	TRICHOMONIASIS	CHLAMYDIA INFECTION	0.020	0.493	0.981	-0.060	0.010	2.098	1.509
17	MUMPS	HEPATITIS B	0.015	0.497	0.985	-0.045	0.008	2.129	1.523
18	MUMPS	MEASLES	0.016	0.539	0.987	-0.044	0.014	6.065	1.978
19	CHLAMYDIA INFECTION	GONORRHEA	0.142	0.604	0.925	-0.328	0.094	2.932	2.007
20	GONORRHEA	CHLAMYDIA INFECTION	0.142	0.689	0.947	-0.270	0.094	2.932	2.460
21	AFB UNDETERMINED	MYCOBACTERIUM NON-TB	0.011	0.764	0.997	-0.018	0.011	31.475	4.143
22	TRACHOMA	CHLAMYDIA INFECTION	0.014	0.881	0.998	-0.018	0.010	3.749	6.436

Clustering Analysis

We found several cluster containing close relationships between diseases and terminologies, such as {Biliary Sludge, HFA, Macrocytosis, Paroxysmal, Pseudogout, back discomfort, betimol, hesitancy, Intron A}, {Gastric Polyps, Kidney failure, antral, benefix, benicar}, {Duodenal Ulcer, Helicobacter Pylori, Malabsorpition, amylase, antimetics} and {appetite lost, immunoglobulin, retrovir} , etc. Some clusters highly correspond to specific diseases or medical processes. For example, appetite lost, immunoglobulin, retrovir are HIV related symptoms. Meanwhile, some clusters contain diseases and terms associated with several medical processes. For example, we found a cluster including gastrointestinal terms (Duodenal ulcer, Helicobacter pylori, Malabsorption, acyclovir, amylase and antiemetics), additive behavior (drinking, lortab andmarijuana) and cancer (methotrexate, vincristine and zofran). This cluster may suggest negative impact of addictive behavior toward digestive system. The appearance of cancer drugs in this cluster could raise a research question about the impact of additive behavior toward the metabolism process, which will further affect the cancer drug efficiency.

Sequential Patterns

We found 105 disease-associations satisfying all 3 criteria about rule_confidence, begin_before_end and coverage to construct the frequent sequential disease patterns. We found 3 groups of sequences in the NCD data. The first group contains only one sequence Hyperplenism → Annemia. The second group contains 5 diseases: Fibrosis Pulmonary, Staphylococcus Methicillin-resistant, Biliary Stricture, Cycsticercosis and Meconium Ileus, in which Fibrosis pulmonary and Meconium ileus stay at the triggering position. This disease group may raise additional research questions since these diseases occur at different organs. The last group is marked by Hepatitis A and 56 other diseases staying at the triggering position of Hepatitis A.

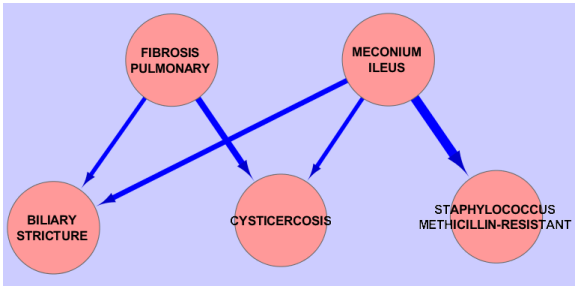


Figure 6. Fibrosis/Meconium-ileus sequence

Visualization

We have developed a Health Terrain Visualization system to visualize interesting knowledge associations that emanate from the data mining and text mining processes. The visualization environment provides the user with a variety of ways to observe interesting patterns. For example, Figure 7a shows the map of Indiana, structured by county and zipcode, displaying the occurrence of the disease *Staphylococcus Methicillin-resistant* across the state as a heat map. The user will have the ability to use filters to observe patterns of various diseases in a similar manner.

Figure 7a. Heat map of disease

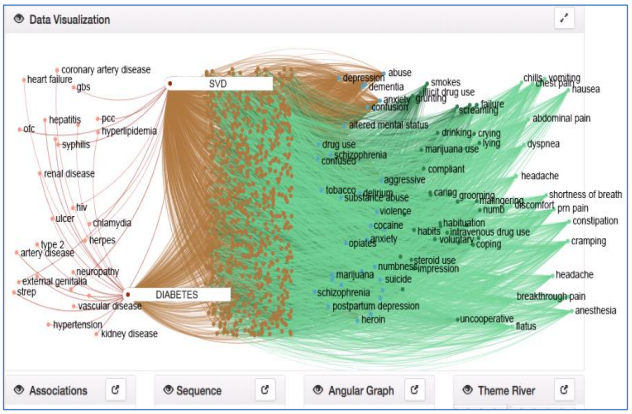


Figure 7b. A multi-feature association graph

In Figure 7b, first loose column of peach colored nodes on the left are the *comorbidities*. Then the *diseases* are highlighted with maroon nodes. The next columns of brown nodes are *patients* followed by the light blue nodes representing *mental behaviors*. The dark green nodes are *risky behaviors*, and the final columns of light green nodes are *symptoms*.

Discussion

NLP provides a means to increase the amount of information that is present in the NCD data. Further analysis is being carried out based on location information that is present in the NCD dataset. This kind of location-based-mining is very useful for public health practitioners in understanding the nature of a disease. To achieve this, we are mining the NCD data and the corresponding clinical data simultaneously, based on zip codes, counties and cities. This work is being incorporated into our Health terrain visualization system.

In this study, by finding associations and grouping strongly-related terminologies into clusters, data mining is useful in guiding visualization application to adopt better visualization layout and highlighting significant and useful information to the users. We also design a visualization-oriented database model to reduce the heavy data-fetching and computation workload on the visualization application.

In addition, we prove that data mining could discover unreported/ill-reported associations between various terminologies in health data science, such as medication-symptom, disease comorbidity, etc. We are working with health professionals to validate and explain the new-found associations. On the other hand, this fact may open further collaboration between computational approaches and traditional biological-medical ontology approach to achieve better understanding on the mechanism of the development and spread of diseases. Another promising direction is integrating EHR data mining with genotype information to construct in-depth knowledge about disease and drug mechanism and visualize this integration by GeneTerrain[8].

Acknowledgement
The project is supported by Department of the Defense; award number W81CWH-13-1-0020. We gratefully acknowledge the contribution of Jennifer Williams; Regenstrief Institute, Project Manager.

References

1. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/
2. Fighting disease outbreaks with two-way health information exchange, last retrieved from http://newsinfo.iu.edu/news/page/normal/11948.html
3. B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–11
4. Chapman , W.; Bridewell , ; Hanbury , ; Cooper , G. F.; Buchanan , G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 2001, 34 (5), 301–310.
5. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/
6. Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342
7. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. Introduction to Algorithm. Massachusetts Institute of Technology Press, 2009, pp 390-396
8. You Qian, Shiaofen Fang, and Jake Y. Chen. GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks (2008) Information Visualization, doi: 10.1057/palgrave.ivs.9500169. manuscript.
9. Mohamemed J. Zaki and Wagner Meria. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press 2014, Great Britain. pp 336-337
10. Van Mechelen I, Bock HH, De Boeck P. Two-mode clustering methods:a structured overview. Statistical Methods in Medical Research 13 (5): 363–94, 2004

Data Exploration of a Notifiable Condition Detector System

Yuni Xia, PhD[1], Shiaofen Fang, PhD[1], Mathew Palakal, PhD[2], Roland Gamache Jr, PhD[2], Thanh Minh Nguyen[1] , Sam Bloomquist[1], Anand Krishnan[2], Jeremy Keiper[3], Shaun Grannis, MD[3]
1 Department of Computer Science, Indiana University – Purdue University, Indianapolis;
2 School of Informatics and Computing, Indiana University – Purdue University, Indianapolis;
3 Regenstrief Institute, Indianapolis, IN

Abstract

The Notifiable Condition Detector (NCD) system is an automated electronic lab reporting (ELR) and case-notification system developed by Regenstrief Institute. It has been used in Indiana for over ten years to report laboratory results for the detection of notifiable conditions such as novel H1N1 influenza, sexually transmitted diseases, lead poisoning, and salmonella [1]. In this paper, we discuss ongoing efforts to analyze and visualize dimensions of the NCD data. We identify most common conditions, describe the distribution of the diseases across gender and race, study the co-occurrence of diseases and find the association rules among different diseases.

Introduction

The Regenstrief Institute implemented and has maintained an HIE-based, automated electronic lab reporting (ELR) and case-notification system for over ten years in Indiana. The Notifiable Condition Detector (NCD) uses a standards-based messaging and vocabulary infrastructure that includes Health Level Seven (HL7) and Logical Observation Identifiers Names and Codes (LOINC). The NCD receives real-time HL7 version 2 clinical transactions daily, including diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations [1]. The NCD automatically detects positive cases of pre-specified conditions and forwards alerts to local and state health departments for review and possible follow up. These alerts enable public health to conduct more effective and efficient population health monitoring.

The initial analytic dataset contained 833,710 notifiable cases from 543,209 distinct patients. The dataset was deidentified by removing HIPAA identifiers and replacing patient age and zip code columns with pseudonymized values. The dataset contains 22 columns in total, including patient pseudo ID, condition name, test result name, test result value, test normal range, race, and gender, among others. The missing data rate for columns varied substantially from 0% for column Patient Pseudo_ID to over 70% for column Test_Abnormal_Flag.

Data Analysis

We analyzed the condition distribution across genders and and the result is shown in figure 1.

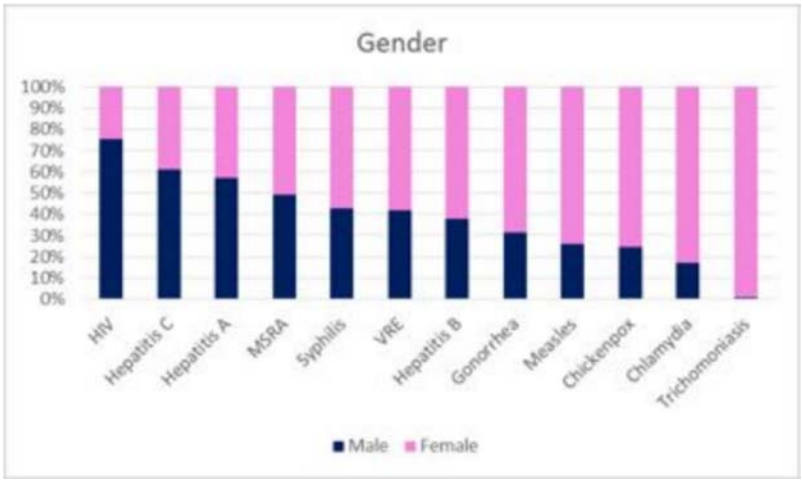


Figure 1: Condition distribution across gender

The conditions in the figure are listed according to male proportion in descending order. It shows that HIV and Hepatitis C are more common in men than in women, while Trichmoniasis , Chlamydia, Chickenpox, Measles, Gonorrhea, and Hepatitis B are more common in women than in men. The conditions that do not show significant gender difference include Hepatitis A, MRSA, Syphilis and VRE.

Disease Association Mining

In the NCD dataset, there are 39,771 patients with two or more diseases. We studied the co-occurrence pattern and derived the association rules between diseases. We generated a graph of associations among conditions using RapidMiner version 5.3, illustrated in Figure 2. It shows 2 clusters among the conditions. One cluster includes Chlamydia/Trachoma, Gonorrhea and Trichomoniasis. The other larger cluster includes Hepatitis A, Hepatitis B, Hepatitis C, HIV, Syphilis, Measles, Mumps, and Chickenpox. The figure also suggests an association between MRSA and VRE. The reason behind the associations has yet to be studied with domain experts.

Figure 2: Diseases Associations

Table 1 shows a portion of the association rules in tabular format. The premise and conclusion of the rule is shown in the table, with the quality measures of each rule including the Support, Confidence, Laplace, Gain, p-s, Lift and Conviction. The rules are sorted according to the Confidence in ascending order. The higher the Confidence, the stronger the association is. A complete explanation of all the quality measures can be found in [2]. We are working with domain experts to evaluate the association rules and tuning the parameters to produce optimum results.

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
5	SYPHILIS	HUMAN IMMUNODEFICIENCY VIRUS	0.010	0.286	0.976	-0.060	0.007	3.468	1.285
6	HEPATITIS A	HEPATITIS C	0.028	0.294	0.939	-0.162	0.017	2.628	1.258
7	HEPATITIS A	HEPATITIS B	0.032	0.332	0.942	-0.159	0.009	1.423	1.148
8	MUMPS	CHICKENPOX	0.010	0.348	0.981	-0.050	0.008	4.377	1.413
9	MEASLES	CHICKENPOX	0.033	0.366	0.948	-0.145	0.026	4.628	1.457
10	CHICKENPOX	HEPATITIS B	0.029	0.370	0.964	-0.130	0.011	1.589	1.218
11	MEASLES	HEPATITIS B	0.036	0.403	0.951	-0.142	0.015	1.726	1.284
12	HEPATITIS C	HEPATITIS B	0.045	0.404	0.940	-0.179	0.019	1.732	1.287
13	CHICKENPOX	MEASLES	0.033	0.412	0.957	-0.126	0.026	4.628	1.548
14	ENTEROCOCCUS VANCOMYCIN-RESISTANT	STAPHYLOCOCCUS METHICILLIN-RESISTANT	0.019	0.442	0.977	-0.067	0.014	3.847	1.586
15	MYCOBACTERIUM NON-TB	AFB UNDETERMINED	0.011	0.450	0.997	-0.038	0.011	31.475	1.783
16	TRICHOMONIASIS	CHLAMYDIA INFECTION	0.020	0.493	0.981	-0.060	0.010	2.098	1.509
17	MUMPS	HEPATITIS B	0.015	0.497	0.985	-0.045	0.008	2.128	1.523
18	MUMPS	MEASLES	0.016	0.539	0.987	-0.044	0.014	6.065	1.978
19	CHLAMYDIA INFECTION	GONORRHEA	0.142	0.604	0.925	-0.328	0.094	2.932	2.007
20	GONORRHEA	CHLAMYDIA INFECTION	0.142	0.689	0.947	-0.270	0.094	2.932	2.480
21	AFB UNDETERMINED	MYCOBACTERIUM NON-TB	0.011	0.764	0.997	-0.018	0.011	31.475	4.143
22	TRACHOMA	CHLAMYDIA INFECTION	0.014	0.881	0.998	-0.018	0.010	3.749	6.436

Table 1: Association rules among conditions

Acknowledgement

The project is supported by Department of the Army; award number W81CWH-13-1-0020.

References

1. Automated Electronic Lab Reporting and Case Notification, http://www.regenstrief.org/cbmi/areas-excellence/public-health/.

Detecting Comorbidity of Chlamydia from Clinical Reports

Mathew Palakal, PhD[1], Shiaofen Fang, PhD[2], Yuni Xia, PhD[2] , Shaun Grannis, MD[3], Roland Gamache Jr, PhD[2], Thanh Minh Nguyen[2] , Sam Bloomquist[2], Anand Krishnan[1], Jeremy Keiper[3]
[1]School of Informatics & Computing, Indiana University – Purdue University, Indianapolis;
[2]Department of Computer Science, Indiana University – Purdue University, Indianapolis;
[3]Regenstrief Institute, Indianapolis, IN

Abstract
Using a standards-based messaging and vocabulary infrastructure, the Regenstrief Institute implemented and has maintained an unparalleled automated electronic laboratory reporting and noticeable condition detection (NCD) system for over 11 years [1]. The NCD automatically detects positive cases of pre-specified conditions and forwards alerts to local and state health departments for review and possible follow up. In this paper, we discuss ongoing efforts to analyze the clinical reports of one specific NCD condition, Chlamydia. Our goal is to identify the presence of any comorbidities of Chlamydia across 6238 patient records and integrate this finding along with our health analytics and visualization system that we are developing.

Introduction
The participants of this project are working on a Health Terrain visualization system centered on an innovative concept-based knowledge discovery and visualization. In this approach, raw health data are first processed, mined and transformed to an information-rich Health Concept Space, where attribute values, association relationships and other partial knowledge are extracted from patient data to form a more structured multi-dimensional data space. The concept space is built on a controlled vocabulary (concepts) that can be pre-defined based on application needs, and therefore is a scalable framework that can be expanded progressively as the applications and use cases expand. For building the concept space, we utilize the structured data as well as unstructured clinical reports from the EHR data. One piece of knowledge that can be extracted from the clinical notes is comorbidity and this can be discovered using Natural language processing (NLP). The text in medical records (e.g. radiology reports, pathology reports, clinical notes, and discharge summaries) includes a wealth of information about patients. NLP can be very useful in extracting information from these free text documents and creating structured information that can be used for further knowledge extraction.

Data
The dataset consists of 6238 de-identified clinical notes that include discharge summary, laboratory reports, etc. Since the clinical records are de-identified, the patient specific information is lacking in the clinical reports.

Method
Natural language processing (NLP) techniques are carried to detect comorbidities that co-occur with chlamydia. In this case, the NLP process is composed of low-level and high level task. Low level tasks consist of sentence splitting, tokenization, stemming, part of speech (POS) tagging and phrase chunking (identifying phrases from POS tagged tokens) and the higher level tasks consists of named entity recognition (NER) of co-occurring diseases. Once the NER process is complete, we use the *tf-idf* vector space model [1] (term frequency * inverse document frequency model) to identify significant co-occurring diseases along with chlamydia. The *tf-idf* is recognized as one of the more effective text mining models compared to Log Level Likelihood and Odds ratio models, amongst the various other models. The *tf-idf* model uses the concept of relevance and co-occurrence of terms. The relevance of a term *j* w.r.t. a document *i* is given as,

$$w_{ij} = t_{ij} * \lg\left(\frac{N}{N_j}\right) \quad (1)$$

w_{ij} = relevance of term *j* in the patient record *i*; T_{ij} = term frequency of term *j* in in the patient record *i*; N_j = frequency of records for term *j*; N = total number of records (N=6238). A particular term is more relevant w.r.t. a record if it appears more frequently in the record and appears in fewer numbers of records in the total records set. An association weight is attached with every association between a pair of terms [3]. This is given by A_{jk}

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg\left(\frac{N}{N_j}\right) * t_{ik} * \lg\left(\frac{N}{N_k}\right) \quad (2)$$

This is essentially a product of the relevance of each of the pair of terms over the entire records set *N*. If the terms do not co-occur in any of the *N* records, then the association is 0. We will be interested mainly in non-zero associations.

Results

After applying the low-level NLP steps, the resulting terms were subjected to named entity recognition (NER). The UMLS [4] database was used for NER to identify diseases that are present in the 6238 discharge summaries. A total of 1337 possible disease conditions were identified that can be roughly considered as comorbidities with Chlamydia. Figure 1 show the most commonly occurring diseases along with the number of reports in which they occur. The co-occurring diseases with Chlamydia were further analyzed using the *tf-idf* model to understand the significance of these diseases across all the 6238 records. For this analysis, *tf-idf* score was calculated for the diseases using the Equation 1.

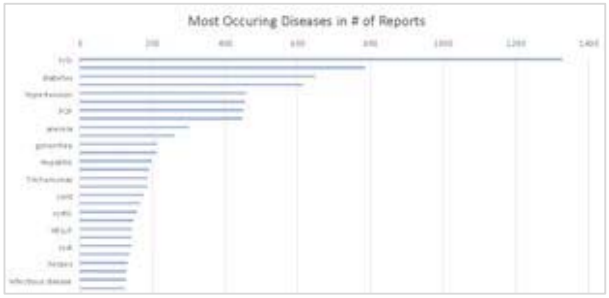


Figure 1: Top diseases with the corresponding number of reports

Comorbidity with Chlamydia

The final analysis and the goal of this work are to identify comorbidity with Chlamydia. To accomplish this, we calculated the pair-wise significance of each disease with Chlamydia using Equation 2. Figure 2 shows all the diseases those are found to be comorbid with Chlamydia. This graph shows some of the expected diseases such as hepatitis, gonorrhea, etc., along few other health conditions such as Bacterial vaginosis, obesity, and so on. The accuracy of these findings has to be validated by the experts. In addition to the comorbidity, the NLP analysis also revealed symptoms, risky behaviors, and mental behaviors those are associated with chlamydia as shown in Figure 3.

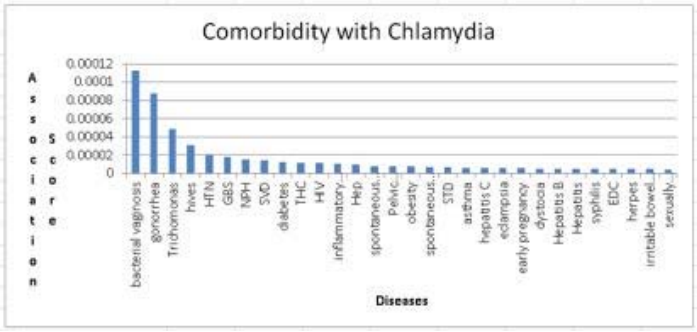


Figure 2: Top scoring comorbidity with Chlamydia

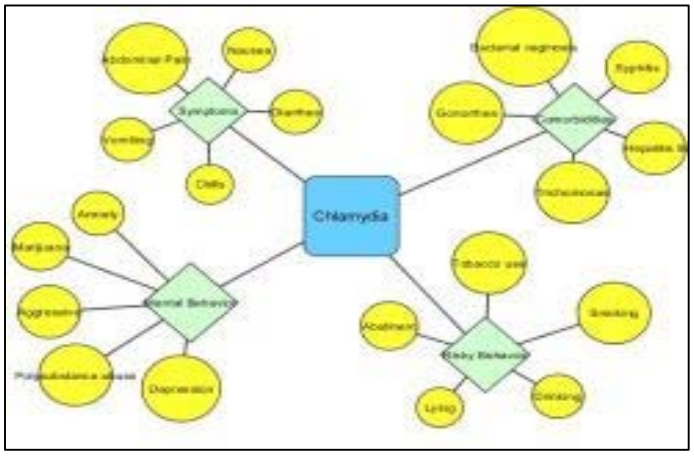


Figure 3: Symptoms and risky behaviors associated with Chlamydia detected using NLP

Conclusions

We have developed an NLP method to identify comorbidity of Chlamydia from the discharge summary of 6238 clinical records. This information, once validated by the experts, will be incorporated in to our Health Terrain visualization system that is currently under development.

Acknowledgement

The project is supported by Department of the Army, award number W81CWH-13-1-0020.

References

1. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/
2. G. Salton. Introduction to modern information retrieval. McGraw-Hill, New York, 1983
3. Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342
4. B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–1

Use Cases for Public Health Data Visualization

Jeremy Keiper[1], Yuni Xia, PhD[1], Shiaofen Fang, PhD[1], Mathew Palakal, PhD[2], Shaun Grannis, MD[3], Roland Gamache Jr, PhD[2], Thanh Minh Nguyen[1], Sam Bloomquist[1], Anand Krishnan[2]
1 Department of Computer Science, Indiana University – Purdue University, Indianapolis;
2 School of Informatics, Indiana University – Purdue University, Indianapolis;
3 Regenstrief Institute, Indianapolis, IN

Abstract

Epidemiologists in Marion County, Indiana, use massive amounts of public health data to provide insight to previous and current disease outbreaks. Typically, they use statistical analysis and simple charts to convey information in public reports, but the work is tedious and difficult without contextual knowledge from years of investigation. Use cases highlighting potential disease outbreaks and research topics, provided by these epidemiologists, will inform a new interactive health data visualization system designed to facilitate and enhance the discovery and strategic intervention.

Introduction

Epidemiologists and other public health officials need to contain the spread, and identify and reduce exposure to disease. We intend to provide a visualization engine to facilitate discovery of these outbreaks. The following factors contribute to a useful public health dataset, and can be used as variables in visualizations:

- Age
- Gender
- Geographic Location
- Race
- Social Networks

Existing visualizations are limited to heat maps or simple overlays, typically simplified to charts and tables for communicating the most essential data to describe the spread or impact of a disease. A useful interface will allow for rendering these reduced views for use in reports. Data analytics can provide insight to what might be the most interesting data, with the final decision of relevance in the hands of the user.

Interviews with Dr. Joseph Gibson, MPH, PhD, Director of Epidemiology at the Center for Urban Health in the Marion County Public Health Department, and Dr. Roland Gamache, PhD, Affiliate Research Scientist with Regenstrief Institute, inform the use cases in this presentation. Some use cases are specific to Indiana or the Indianapolis area.

Use Case: Histoplasmosis

A rare disease, affecting only 20-25 people yearly in Marion county, histoplasmosis outbreaks should be easy to detect locally. Excavation that disturbs ground where mold spores have settled typically predicates an outbreak of this disease. Indiana has a long history with histoplasmosis due to a high concentration of relevant spores throughout the state; the largest outbreak in the country occurred in Indianapolis in September 1978 and again in August 1979 [1]. Public health officials need to see each diagnosis, with multiple occurrences in a similar location prioritized over others, alongside relevant excavation data. Visualization techniques can highlight geographic proximity to potential causes.

Major Visualization Factors	Visualization Traversal Vectors
Occurrence density in geographic locations	Chronological time
Nearby potential causes (e.g. construction, tree removal, or demolition)	Proximity to causal events

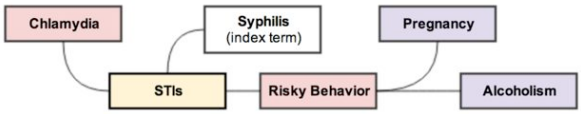
Use Case: Gonorrhea

When researching gonorrhea, epidemiologists are interested in co-morbities and other related diseases. They also want to know who is repeatedly infected after treatment. Gonorrhea is a social disease, transmitted sexually, and a proper visualization will highlight these connections. Social networks, both online (e.g. Facebook and Twitter) and traditional (e.g. high schools, colleges, neighborhoods, and workplaces), become the most important correlation factor. Intervening with super-spreaders, people who may be unknowingly giving gonorrhea or related diseases to several others, can help contain the impact and prevent recurrence. Social data may require approval by individuals, but would facilitate better care.

Major Visualization Factors	Visualization Traversal Vectors
Frequency of recurrence in individuals	Chronological time
Comorbidities (e.g. other sexually transmitted diseases)	Social distance from a super-spreader

Use Case: Risky Health Behavior

Some diseases are indicative of risky health behavior, and epidemiologists need to know when an individual has a higher potential for infecting others in the community. Visualizing layers of an index term against related diseases and risky behavior can show unexpected correlations. The example below shows first, second, and third relations.



Major Visualization Factors	Visualization Traversal Vectors
Cumulative occurrences of risky behavior	Time deltas (e.g. daily, one week, two weeks, months)
Time range between occurrence (e.g. a spike on multiple short deltas)	Age ranges
	Relatedness of other diseases or categories

Use Case: Influenza Pandemic

Influenza is the type of disease that appears seasonally, as a response to airborne allergens. Strategies for a breakout of influenza vary depending on the current state of the pandemic. When integrated with a monitoring system, a visualization engine could highlight a grouping of recent diagnoses and assist in intervention and containment. If the pandemic has already breached a certain threshold, the system would switch to a different visualization useful for monitoring and forecasting spread. It is important in such a breakout to identify the first cases, to understand how the disease is progressing and spreading from the index patients. This information should be presented visually to help decide the best strategy.

Major Visualization Factors	Visualization Traversal Vectors
Earlier occurrences highlighted for easy identification	Amount of time since infection
Occurrence density in geographic locations	Social or geographic location links

Future Research

Each use case above demonstrates problems in public healthcare that cannot be easily interpreted without a series of statistical charts and tables, and the knowledge to recognize trends and similarities. We will use these scenarios as starting points to inform the design of a visualization engine with an interactive interface, primarily useful to the epidemiologists. We expect this system to make the discovery and research of outbreaks and pandemics easier, faster, and more effective.

Acknowledgements

The project is supported by the Department of the Army, award number W81CWH-13-1-0020

Health-Terrain: Visualizing Large Scale Health Data

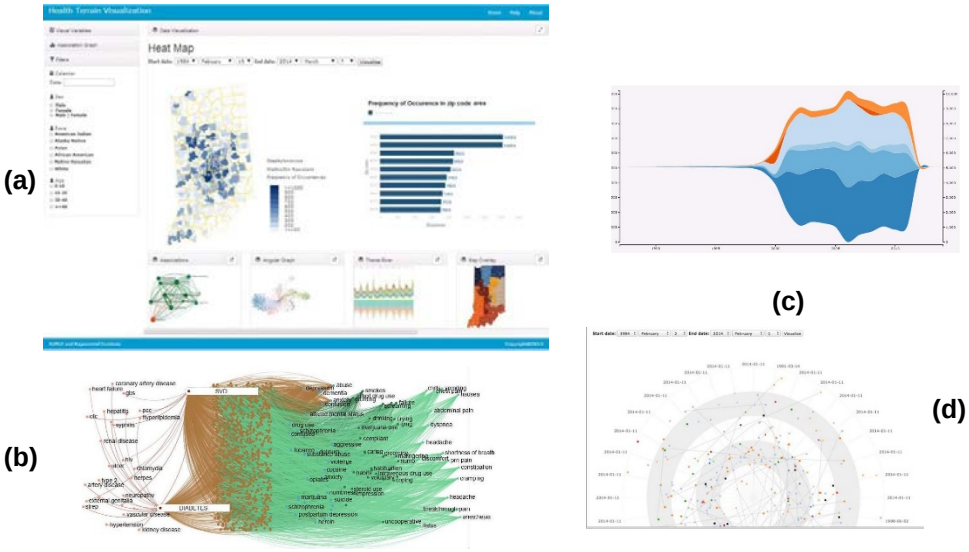
PI: Shiaofen Fang **Org:** Indiana University **Requested Amount: $670,646**

Study/Product Aim(s)

- Build a framework for concept space for pilot healthcare applications.
- Extract attributes of concepts from raw health data using data mining and knowledge extraction.
- Develop and implement visualization algorithms for Health-Terrain visualization, and develop a prototype system.
- Develop application case studies using the prototype system.

Approach and Military Relevance

Raw health data will first be transformed to an information-rich Concept Space, where attribute values and association relations are extracted from patient data, and visualized using Health-Terrains. Health-Terrain is a terrain surface based scalable visual representation for large multi-dimensional data. It is an ideal approach for military EHRS as it is very effective in revealing trends, patterns, and abnormalities with simple heads-up displays to support real time decision making. It also provides a way to unify heterogeneous data into an intuitive and comprehensive visualization to facilitate interoperability among multiple military HER systems. Pilot cases will be studies using the nation's largest and longest tenured health information exchange through Regenstrief Institute.



(a) System interface; (b) Association graph; (c) Theme River; (d) Patient time-line graph.

Timeline and Cost

Activities CY	13	14
Concept space def. (aim 1)		
Algorithm design (aim 1,2)		
System dev. (aim 2,3)		
System Testing (aim 2,3)		
Pilot appl. (aim 4)		
Estimated Budget ($k)	$430	$240

Updated: April 3, 2014

Projected Goals/Milestones:

1. **3/7/2013 – 7/6/2013. Concept Space Definition:** data processing, definition of the concept set, including their attributes and functionalities based on the pilot applications.
2. **6/7/2013 – 10/6/2013. Algorithms Design:** completing the algorithm design phase for visualization and data mining.
3. **8/7//2013 – 5/6/2014. System Development:** finishing software development of the prototype system.
4. **5/7/2014 – 9/6/2014. System Testing:** completing functionality tests for all components of the system.
5. **5/7/2014 – 9/6/2014. Pilot Applications:** Testing the prototype system on pilot applications using the Regenstrief database.